UNITED STATES DEPARTMENT OF AGRICULTURE
Bureau of Agricultural Economics

APPLICATIONS OF A SIMPLIFIED METHOD OF GRAPHIC CURVILINEAR CORRELATION)

By

L. H. Bean, Senior Agricultural Economist
Division of Statistical and Historical Research

A Preliminary Report

Part I

Washington D. C.
April, 1929

382704

APPLICATIONS OF A SIMPLIFIED METHOD OF GRAPHIC CURVILINEAR CORRELATION

By L. H. Bean, Senior Agricultural Economist, Division of Statistical and Historical Research, Bureau of Agricultural Economics

The practicing statistician and economist who is frequently called upon to determine the quantitative relationships between two or more factors often finds it inconvenient or undesirable to apply the formal technique of multiple curvilinear correlation as described in the Journal of the American Statistical Association.1/ Time and clerical help are often lacking or insufficient data do not warrant the use of the formal technique. It is the writer's experience, shared undoubtedly by others who have seriously attempted analysis of time series or other problems involving small numbers of observations, (30 or less), that it is often possible to resort to simplified methods of multiple correlation requiring little time or labor and yielding results of considerable practical value.

The purpose of this paper is to present this simple approach to multiple curvilinear correlation. The method employed will be demonstrated with six examples, or cases, of actual problems chosen from different fields of economic relationships, in the hope that this "case" method of presentation will not only make clear the simple steps involved, but will also suggest their application to similar problems likely to be encountered by the analyst of variations in economic data. It will be demonstrated by means of a generalized problem, but in this final illustration also we shall refrain from generalization. Technical language will be used as little as possible, but the reader will need to study closely the graphic presentations, for the method is essentially one of graphic correlation. In each example the data used and the steps in the analysis will be so indicated that the reader may properly appraise the reasonableness of the approach and the reliability of the results.2/ The assumptions and logic involved in each of the six special cases will also be indicated but this only briefly, for we are concerned here primarily with the simple method of correlating certain factors and not with the reasons for selecting the factors used.

To the reader acquainted with the formal method of multiple curvilinear correlation it may be of interest to observe at the outset that the methods used in the examples are not unlike those now in use. The formal method involves (1) multiple linear correlations to determine a first approximation to the net effect of each independent factor on the dependent one, (2) computation of residuals or differences between the values of the independent variable and values estimated from the linear regressions, (3) plotting the residuals as deviations from each of the linear regression curves, and then (4) the reduction of the residuals to a minimum by a process of successive approximations which involves the free hand drawing of curvilinear regression lines. In the approach illustrated here, steps (1), (2), and (3) are not used. In their stead one or more simple scatter diagrams are used, first approximation to curves drawn free hand by inspection, and residuals usually reduced to a minimum by subtracting first the effect of one variable and then of another. The first approximation curves are then ad-

---

1/ See Journal American Statistical Association, Vol. XVIII, No. 144, A Method of Handling Multiple Correlation Problems, by H. R. Tolley and M. J. B. Ezekiel and Vol. XIX, No. 148, A Method of Handling Curvilinear Correlation for any Number of Variables, by M. J. B. Ezekiel.

2/ Each of the analyses of time series contained in this paper was originally based on data to and including 1927 and gave very satisfactory results when the relationships were applied to 1928 conditions.

adjusted, if necessary, with reference to the residual variation until no further changes are indicated. This simple approach will be illustrated by the following six examples.

The first case, or example, involves three variables, one dependent and two independent, where the effect of the second is first removed, and the residuals practically entirely explained by the third. The example deals with potato prices.

The second case also involves three variables and is like case I except that the third variable is first adjusted for trend before it is used to explain the residuals derived from the relationship between the first two variables. The example deals with mill consumption of cotton.

The third case involves four variables, the fourth of which is a composite of "other" factors represented by a regular trend in residuals. This example deals with the cotton consumption data of case II.

The fourth case involves four variables, in which the fourth is a composite of "other" factors represented by an irregular trend in residuals. The example deals with the yearly average price of apples.

The fifth case dealing with orange prices involves five variables, two of which are highly intercorrelated.

In the sixth case, dealing with acreage changes, the simple approach is applied to a correlation problem in three variables, with one dependent variable expressed as relative first differences, or percentage changes from one period to the next.

In the seventh case, the simple method is applied to a general problem of 4 variables, 30 observations, and high inter-correlations between each of the variables.

After these cases have been presented, something will be said in conclusion concerning the limitations of the methods used here, the results compared with those obtainable by the methods now in use, and the significance that may be attached to results from the analysis of relatively short time series.

CASE I

Relation of (1) production of early potatoes and (2) the price of old potatoes to (3) the price received by producers of early potatoes.

This case deals with three variables, the first and second of which are almost perfectly correlated with the third. It illustrates the method of determining by inspection the net relation between the first variable and the third (dependent) and then the relation between the second variable and the residual fluctuations in the dependent not explained by the first.

The data and the steps involved in this analysis are all contained in Figure 1. Disregarding for the moment the solid curve in section I, we have here a scatter diagram in which the production of early potatoes for the period 1921-1928 is plotted against the price received by producers. These prices are shown in section 4. The price of old potatoes is plotted in section 3.

THE EFFECT OF EARLY POTATO PRODUCTION AND THE PRICE OF OLD POTATOES ON THE GROWERS PRICE OF EARLY POTATOES

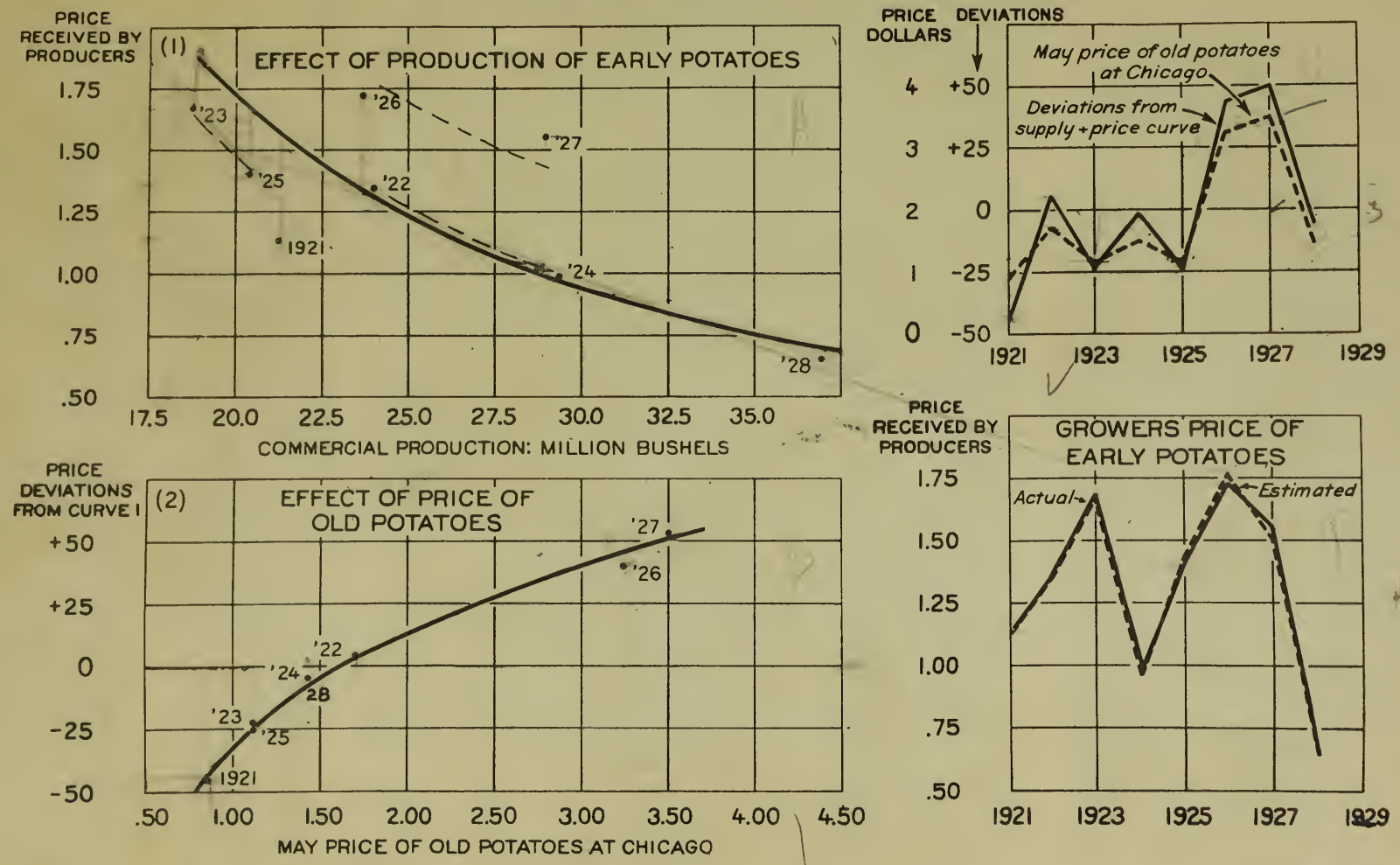


Figure 1

Starting with these two price series and the scatter diagram, our aim is to establish first the effect of production on the price received by producers. In other words we wish to draw a curve through the observations in section I, which will represent the net effect of production alone, with the effect of the price of old potatoes held constant.

By inspecting the movements of prices of old potatoes, it is observed that the prices in 1923 and 1925 were the same or nearly so, that is, in these two years the effect of old potato prices may be assumed to have been equal or constant, so that the difference in the prices received by growers of early potatoes may be tentatively assumed to be due to variations in production. By connecting the observations in the scatter diagram, the apparent effect of production in these two years of constant old potato prices is indicated. We may now make the further observation that prices of old potatoes in 1922, 1924, and 1928 were about the same, suggesting that their effect on the prices received was probably constant in these three years. By drawing a curve which will pass through the three corresponding observations in the scatter diagram, we obtain the effect of production independent of old potato prices in these three years.

We may now make a final observation that old potato prices were high in 1926 and 1927 and low in 1921, 1923, and 1925, which suggests that a curve may be so drawn through the observations in the diagram which will leave the 1921, 1923, and 1925 observations below the curve and those of 1926 and 1927 above, and parallel the previous short segments. (The curve shown in the diagram was so drawn.) This gives us a tentative indication of the net effect of production on the price received.

Our next step is to see to what extent the deviations from this average supply and price curve can be explained by the fluctuations in the price of old potatoes. These deviations, which are the portions of the price not explained by production, may be measured or read graphically directly from the diagram, and are shown plotted in section 3. The quantitative relationship between the price of old potatoes and the deviations from the supply and price curve is shown in section 2 where the May prices of old potatoes are plotted against the deviations. By slight adjustments in the preliminary supply and price curve it becomes evident that the observations in the scatter diagram of section 2 lie along a curve which may be drawn in free hand. Inasmuch as there are only minor deviations from this second curve it is clear that these two factors (supply of early potatoes and price of old potatoes) account for practically all of the variations in the price received by producers of early potatoes.

The extent to which that is true can be demonstrated by reading from the supply and price curve for the production of each year the average effect of production on price, and adding to it or subtracting from it the readings from the second curve of the average effect of old potato prices on early. The algebraic sums of these two readings for the corresponding years as well as the actual prices are shown in section 4.

It will be observed that the almost perfect correlation shown in the comparison between the actual and the estimated prices in section 4 was obtained merely by (1) plotting the data used in the analysis, (2) drawing by inspection a free hand net regression curve of supply and price, (3) plotting deviations read from this curve (which deviations may be considered as the original prices with the effect of production eliminated) against the second independent factor (the price of old potatoes) and (4) drawing a free hand

curve through this second scatter diagram. Another step that should be employed unless the residual variation is practically zero, as it is in this case is to plot the final residuals as deviations from the two curves as a final test of goodness of fit.

The other examples which follow are in a large measure only variations from this simple case. They all involve determining curves by inspection, eliminating the effect of one variable from the original dependent, and eliminating the effect of the second and third from the residuals, thus reducing the latter to a minimum. The last step, checking the final slope of the curves with reference to the final residuals, was not deemed necessary in the following illustrations in view of the very small final residuals.

CASE II

Relation of cotton prices and business conditions to the domestic mill consumption of cotton.

This case varies from case I only in that one of the independent variables shows a very definite upward trend, but the same method of correlation may be employed if a simple adjustment for trend is used.

Two main assumptions are involved in this analysis of cotton consumption. First, variations in domestic mill consumption of cotton are due very largely to the price of cotton in relation to the price of cotton goods and to changes in business activity. Low prices due to large cotton crops create favorable price margins for manufacturers. Under such conditions cotton is purchased beyond the current requirements for later consumption; conversely, high cotton prices create unfavorable profit margins, and curtailment of purchases of raw cotton is reflected in subsequent curtailment in cotton mill activity. Thus it appears that variations in price of cotton during a given crop year (July-June) are reflected in mill consumption during the following calendar year (indicating roughly a lag of about six months if semi-annual or monthly data were used.) The second assumption is that cotton mill consumption is also affected by general business conditions which reflect the industrial demand for cotton and the buying power of consumers in their demand for cotton goods.

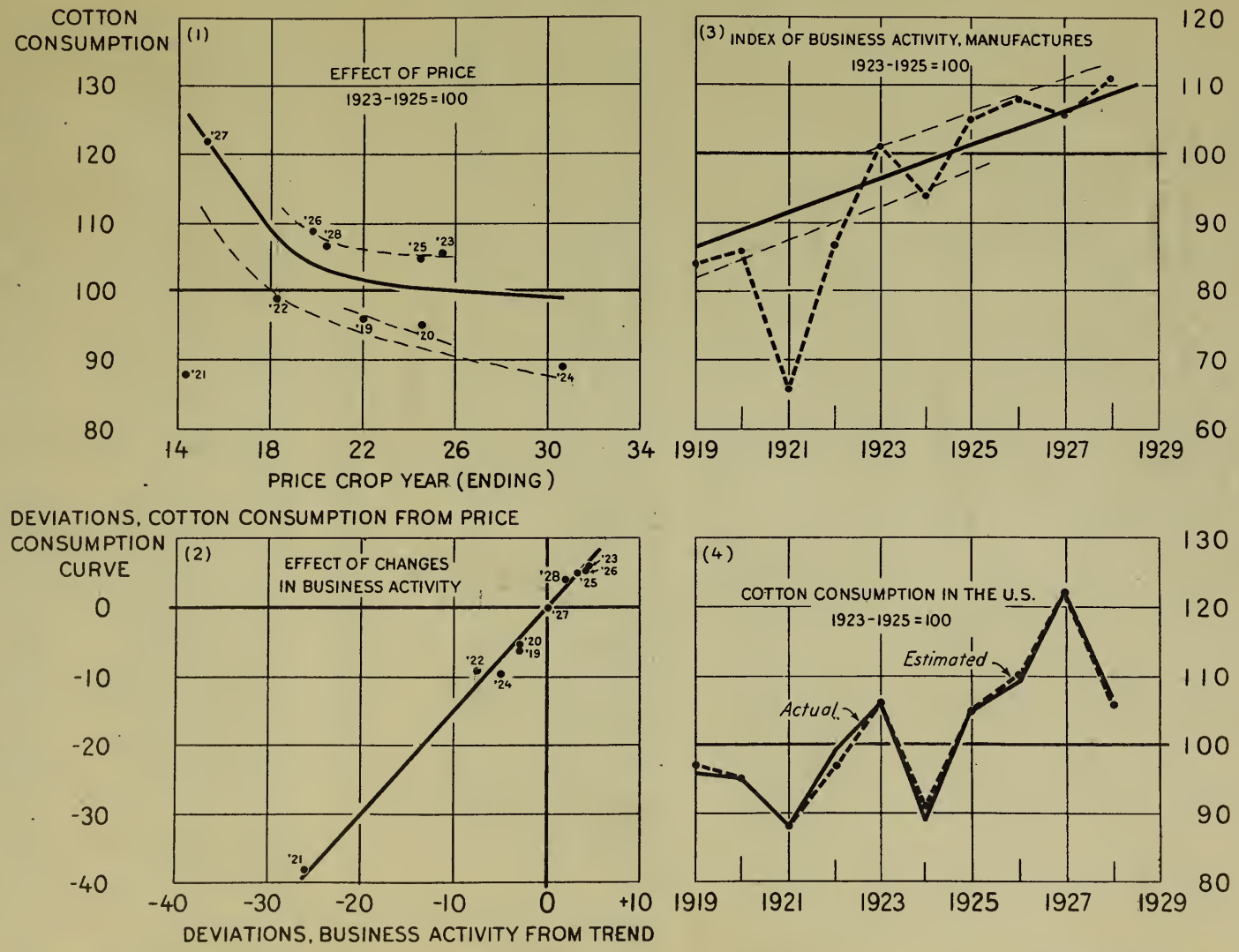
Figure 2 contains the following: In section 4 the index of cotton consumption 1919 to 1928, and in section 3 the general index of production of manufactures, both published by the Federal Reserve Board (1923-25 = 100): in section 1 the New Orleans crop year average price[3/] of middling spot cotton plotted against the index of consumption in the form of a scatter diagram. The index of manufactures is here used as a general measure of business activity. If the curve in the scatter diagram be disregarded for the moment it will be evident that the scatter is wide and that no appreciable correlation between price and consumption is apparent on the surface. However, the nature of the effect of price on consumption becomes evident after a moment's inspection of the index of business activity.

We note first that the index of business activity has an upward trend. This suggests studying the variations in this index above and below a trend and

---

3/ Adjusted for changes in the Bureau of Labor statistics index of wholesale prices, 1926 = 100.

THE EFFECT OF COTTON PRICES AND BUSINESS ACTIVITY ON DOMESTIC COTTON (MILL) CONSUMPTION



U.S. DEPARTMENT OF AGRICULTURE

BUREAU OF AGRICULTURAL ECONOMICS

Figure 2

relating them to the position of the comparable observations in the scatter diagram.4/ Disregarding for the moment the trend lines finally used here, we note that each of the years 1923, 1925, 1926, and 1928 were years of high business activity, as indicated by the dotted upper line. We then observe the location of the corresponding points in the scatter diagram of consumption against price, and draw a line through them. Next we note that the indexes of business activity for the years 1919, 1920, 1922, and 1924 lie on a lower trend line with 1921 considerably lower. Again we find the price consumption points for these years in the lower part of the diagram, with the observation for 1921 and 1924, the years of greatest business depression lower than the rest. Our problem now is to draw a free-hand trend line through the index of business activity and to draw a corresponding curve through the scatter diagram, so that for each of the deviations from the trend line there will be a corresponding deviation from the price-consumption curve. A straight line drawn in, approximately midway between the tentative upper and lower lines through the indexes of business activity, passes through the index for 1927. This suggests that a curve may also be drawn between the upper and lower tentative lines in the scatter diagram, also passing through the "1927" point. Having drawn in these two center lines, their adequacy is tested by plotting the deviations from the trend in business activity against the deviations from the consumption-price curve (see section 2.) After very slight adjustments in the curve and trend line, it is found that the relationship between the deviations are best represented by a straight line.

The completeness with which the curves thus derived (I, the effect of price, and II, the effect of variations in business activity) explain the variations in cotton consumption appears in section 4, where the algebraic sum of the readings or estimates are shown in a broken line.

It is to be observed that in other cases of this sort, an adjustment for a downward trend in one of the variables may be necessary.

CASE III

Cotton Consumption (Continued)

In this third case we have an illustration of a four variable problem to which the fourth variable is a regular trend in residuals after eliminating the effects of two other variables, and may be considered as a composite of "other" factors not included in the problem but related to the progress in time. The analysis used in the second illustration lends itself also as an illustration here if we adopt a slight modification of the foregoing procedure.

Thus, instead of assuming a trend from which to measure deviations in business activity we may relate the actual index to the deviations from the price-consumption curve and obtain a second set of residuals. That procedure gives the scatter diagram in section 5, Figure 3 (in place of section 3, Figure 2.) At first glance it appears that the relationship between the

---

4/ For a criticism of trend elimination, see Journal American Statistical Association, Vol. XX, note on Error in Eliminating Secular Trend and Seasonal Variation Before Correlating Time Series, by Bradford B. Smith.

index and the price-consumption deviations is a curvilinear one, calling for a curve drawn from the lower left corner of the diagram, tapering off into the upper right corner. But by following the observations in time sequence, as indicated by the dotted line beginning with 1919, it becomes obvious that the net effect of business activity on the price-consumption deviations is best represented by a straight line.5/ This straight line, drawn parallel to the lines for successive short periods, has a slope such that a 20 point change in the index of business (from 100 to 80) is accompanied by a change in the index of consumption of 30 points. It may here be observed that this slope, determined independently is identical with the one in Figure 2 where a deviation from trend of 20 points is accompanied by a deviation in the consumption index of 30 points.

If we now proceed to plot observations in section 5, Figure 3 as deviations from curve III, we find that they show a downward trend and fall along a straight line, which is accordingly drawn in as in section 6.

The procedure in this third illustration, it should be clear, was first to eliminate the effect of one variable (price) on the dependent (cotton consumption) by measuring residuals from the price consumption curve determined by inspection. We next eliminated the effect of another variable (business activity) from these first residuals by measuring second residuals from curve III, also determined by inspection, and by considering time as another composite variable it almost completely explains the final residuals in the form of a downward straight line trend. Readings from I, III, and IV summed algebraically give practically the same estimates of consumption as do readings from I and II. Obviously case II is simpler than case III, but the latter is intended mainly as an illustration of successive reduction of residuals to a minimum. In other cases of this type the second set of residuals may fall along a uniform upward trend.6/

For a proper interpretation of the downward trend in residuals obtained in case III we related the variations as well as the growth in business activity to the variations in consumption not explained by price. The downward trend in residuals therefore is due to the upward trend in business activity and reflects in part the fact that the relation of cotton consumption to business activity has been changing with passing years, since business

---

5/ This is obvious from the fact that the straight lines of best fit for any set of three observations are all practically parallel. See, for instance, 1920-21-22; 1922-23-24; 1923-24-25.

6/ For the third method of handling the data in case II as well as for an illustration of an upward trend in residuals, see "Some interrelationships between the supply, price and consumption of cotton" by L. H. Bean, paper read before the New York Section American Statistical Association, April, 1928. Here the index of cotton consumption was first divided by the index of business activity. Consumption adjusted for general business activity was then plotted against price adjusted for the general commodity price level and a downward trend in deviations established from a free hand price-consumption curve. In recomputing or estimating consumption from these factors the sum of readings from the price-consumption curve and the downward trend in residuals were multiplied by the index of business activity.

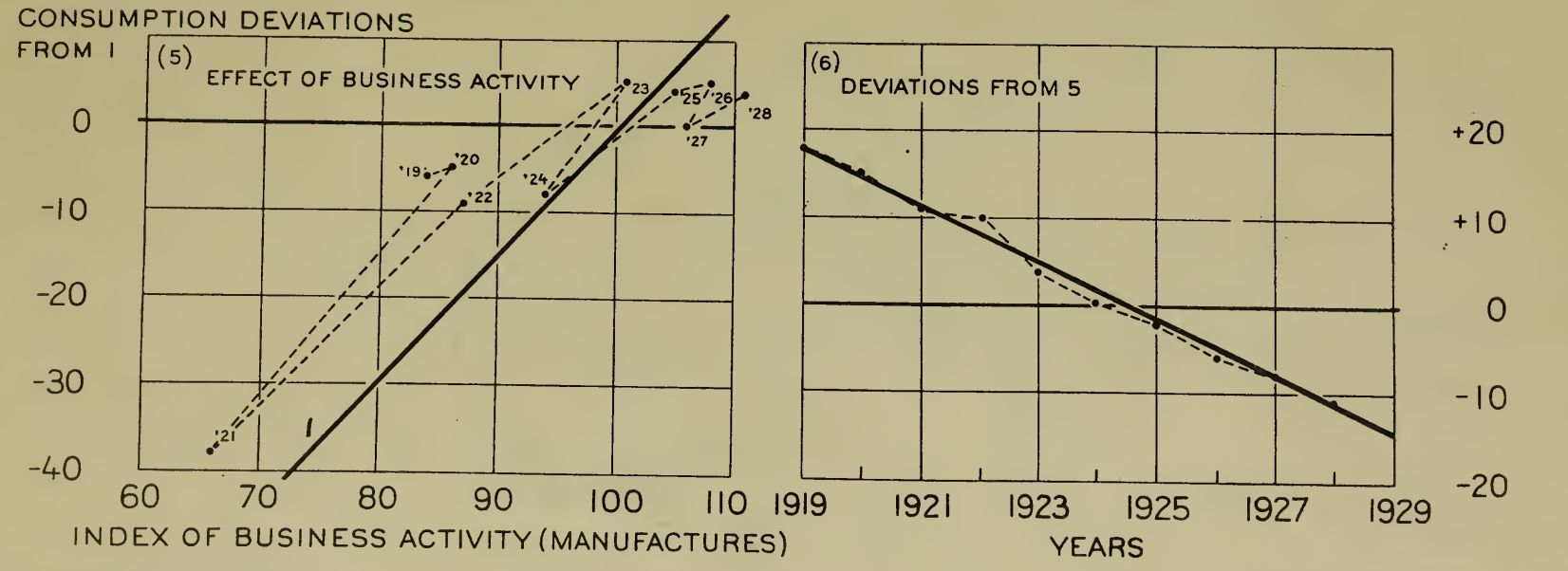
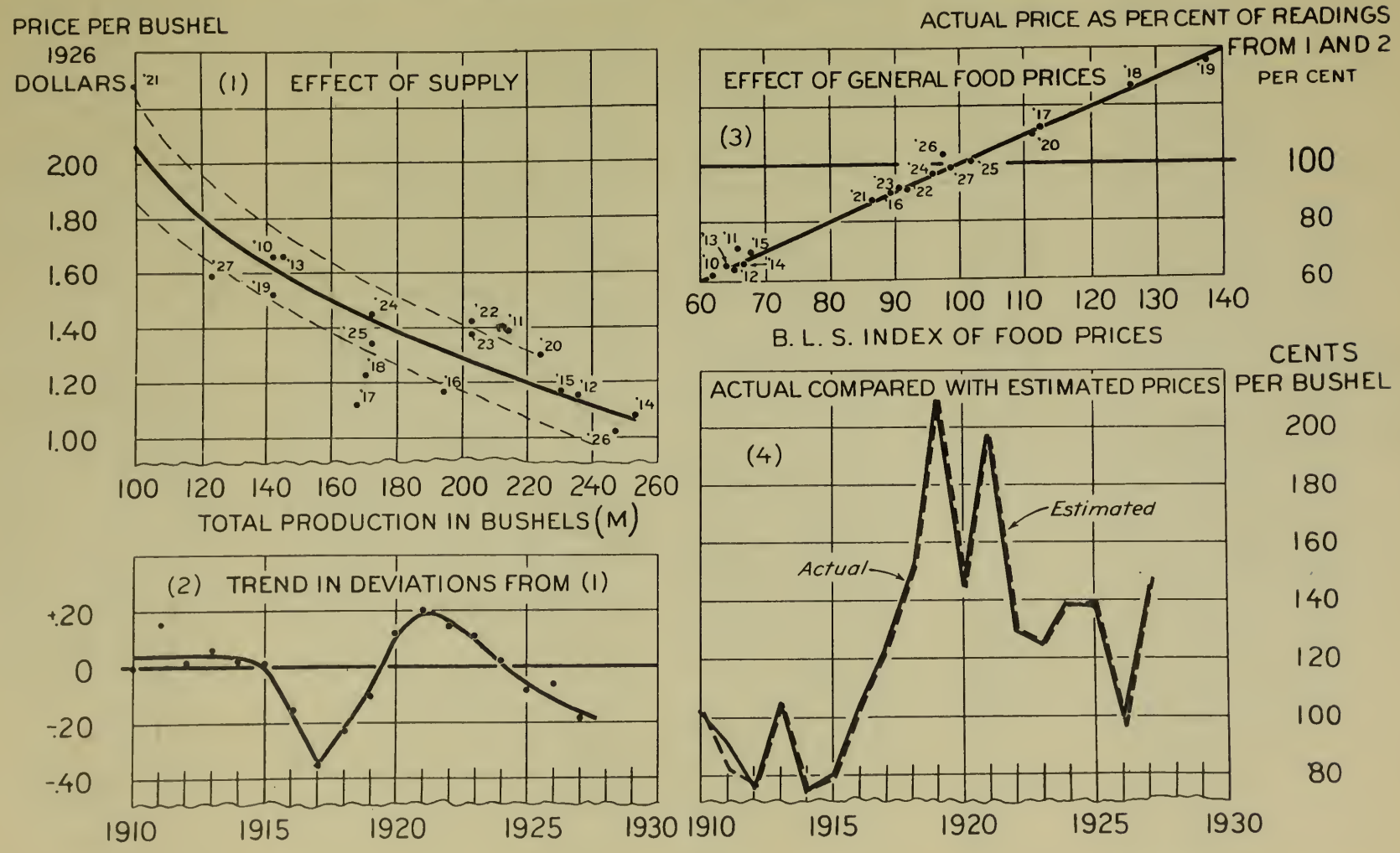

CONSUMPTION DEVIATIONS
FROM I
(5)
EFFECT OF BUSINESS ACTIVITY
0
-10
-20
-30
-40
'19
'20
'21
'22
'23
'24
'25
'26
'27
'28
I
60
70
80
90
100
110
INDEX OF BUSINESS ACTIVITY (MANUFACTURES)
(6)
DEVIATIONS FROM 5
+20
+10
0
-10
-20
1919
1921
1923
1925
1927
1929
YEARS
U.S. DEPARTMENT OF AGRICULTURE
BUREAU OF AGRICULTURAL ECONOMICS

Figure 3

THE EFFECT OF SUPPLY AND OTHER FACTORS ON THE YEARLY AVERAGE FARM PRICE OF APPLES



U.S. DEPARTMENT OF AGRICULTURE

BUREAU OF AGRICULTURAL ECONOMICS

Figure 4

activity in the United States has expanded more repidly than has cotton consumption during the ten-year period 1919-1928.

CASE IV

Effect of supply and other factors on the yearly average farm price of apples

The fourth example illustrates a simple procedure that may be followed in instances where the final residuals follow on irregular trend, instead of a straight upward or downward line. As in the case of a regular straight line trend in residuals, an irregular one may also be considered as the composite effect of "other" factors not included in the analysis, but related to time.

This example deals with the yearly farm price of apples. The independent variables are the total supply of apples in the United States, the general level of food prices at wholesale, and "other" factors represented by the trend in residuals.

The procedure followed in this case is first to "eliminate the effect of the general food price level by dividing the series of apple prices from 1910 to 1927, by a food price index, on the assumption that the price of apples usually fluctuates with the major movements of food prices in general. The next step is to plot in the scatter diagram of section 1, Figure 4, the prices of apples in terms of 1926 food dollars against total production, and by inspection to determine the net effect of total supply on the adjusted price.

This curve, shown in the diagram, is the result of noting that a simple curve "fits" the pre-war observations (excepting 1911 as an unusual year), that a similar curve somewhat higher, fits the observations for 1920-1923, and also (somewhat lower) the observations for 1925-1927. These tentative curves further indicate that an <u>average</u> supply and price curve drawn through the scatter diagram would reveal considerable negative deviations for 1916-1918, positive deviations for 1920-1923, and a downward trend in residuals since 1921. These residuals are shown in section 2, Figure 4, through which an irregular trend has been drawn - a trend which dips down sharply in the war years when the apple export market was completely shut off, rises sharply to 1921 probably as a result of a recovery in foreign demand, and declines since then, which may be attributed to increasing domestic competition from other fruits.

For our present purpose we are not so much concerned with the various factors which may be included as an explanation of these price residuals not accounted for by total supply and the general level of food prices as we are with obtaining the nature of the trend in the composite effect of all "other" factors on the yearly price of apples. This trend, even though it appears irregular when the entire 18-year period, 1910-1927, is considered, indicates sufficient regularity during the last seven years to make an analysis of this sort as useful as those already presented.

Now that we have two curves, one representing the effect of supply and another the effect of all other factors associated with time, (except that represented by general food prices) the third step is to express in the form of a third curve the one-to-one relationship between the index of food

prices and apple prices originally assumed in dividing apple prices by the index. Curves I and II almost completely account for the price of apples in terms of 1926 dollars. Consequently, differences between readings from these two curves and the actual prices in current dollars may be taken as the portions of price originally attributed to the factors represented by the general food price level. The actual prices, when divided by 7/ the sum of readings from curves I and II, may therefore be plotted against the food price index, as in section 3. These observations, excepting two, lie practically along a straight line (as was assumed) which may now be used in conjunction with the other curves to obtain price estimates. In section 4, the actual prices are compared with those obtained by readings from curves I and II, multiplied by readings from curve III.

CASE V

Effect of supply and other factors on the New York price of California oranges

This example illustrates the application of the methods already described to a problem in five variables which involves intercorrelation between independent variables. It deals with the New York price of California oranges as the dependent variable and the total production of oranges, the production of competing fruits, the general level of food prices, and factors related to time, as the independent variables.

The steps involved in this analysis are similar to those already described but an additional one is employed here to eliminate the intercorrelation that exists between the production of oranges and the production of other fruits. They need only brief comment. In the other problems, this intercorrelation was eliminated by selecting observations constant as to one factor, in determining the first approximation to the net regression curve for a second factor.

In section I of Figure 5 the total production of oranges in the United States is plotted against the price of oranges (November-October), adjusted for changes in the index of food prices at wholesale, and the curve representing the effect of production on price drawn in, as suggested by the tentative curves passing through the observations 1920-22, and 1923-27. The location of these two lines indicate an upward trend in deviations from an average curve for the entire period.

In section 4 are plotted the United States production of oranges against an index of production of competing fruits which appear on the market during the crop year for oranges, November-October. The intercorrelation is such that large crops of oranges are usually accompanied by large crops of other fruits in the aggregate, and vice versa; also year to year changes in the orange crop are generally accompanied by similar changes in the composite of competing production.

Before attempting to explain the deviation from I in terms of oranges, because of competing production, it is desired to exclude from the latter the changes in orange production which are already taken into account in section I, that is, the effect of competing products which are attributed to oranges in section I. We may do this by drawing curve IV, the slope of which is suggested by the lines passing through the points, 1921, 1922, 1925, 1926

7/ Division instead of subtraction because the first step was a division of the actual price by the index.

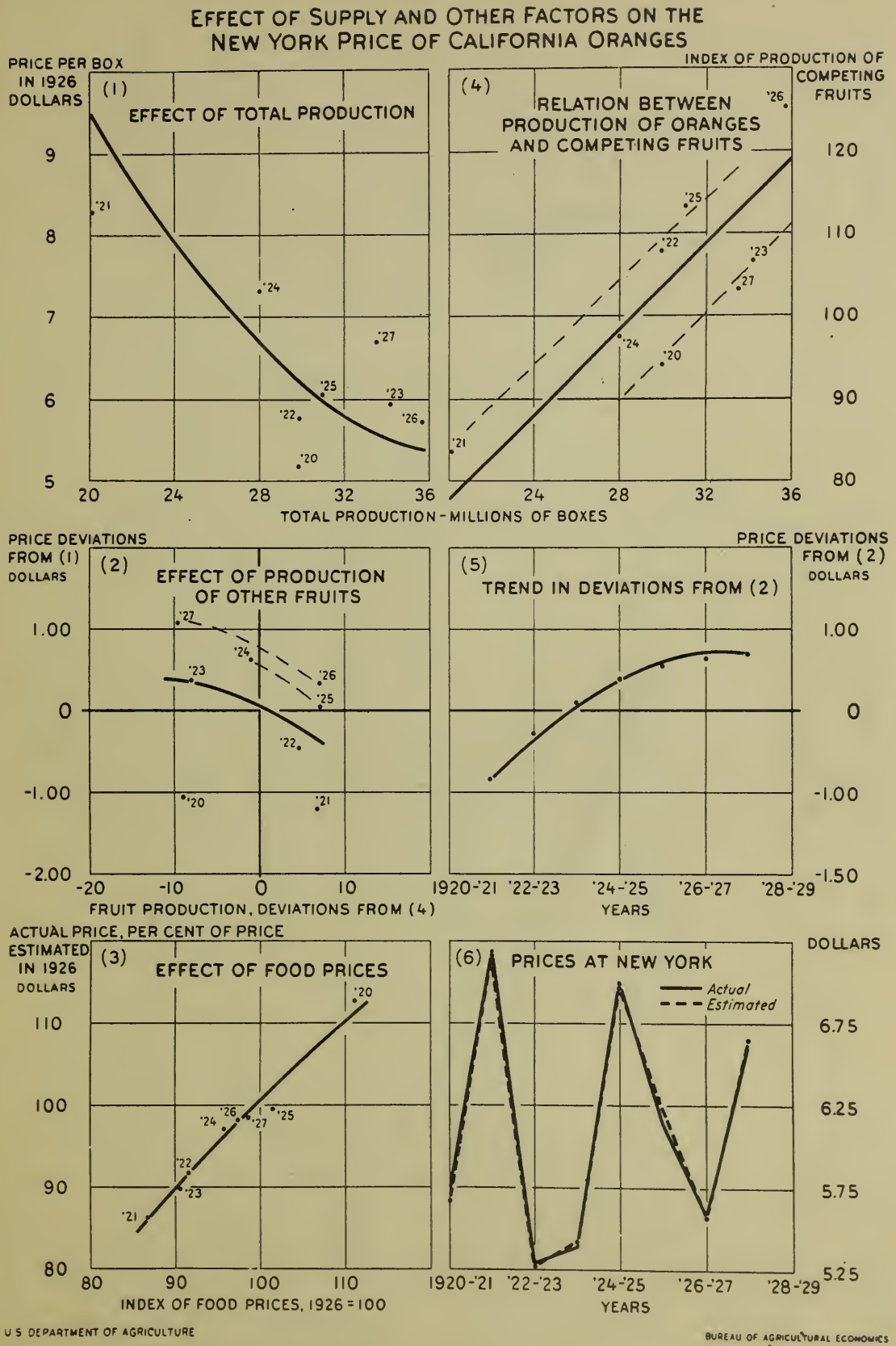
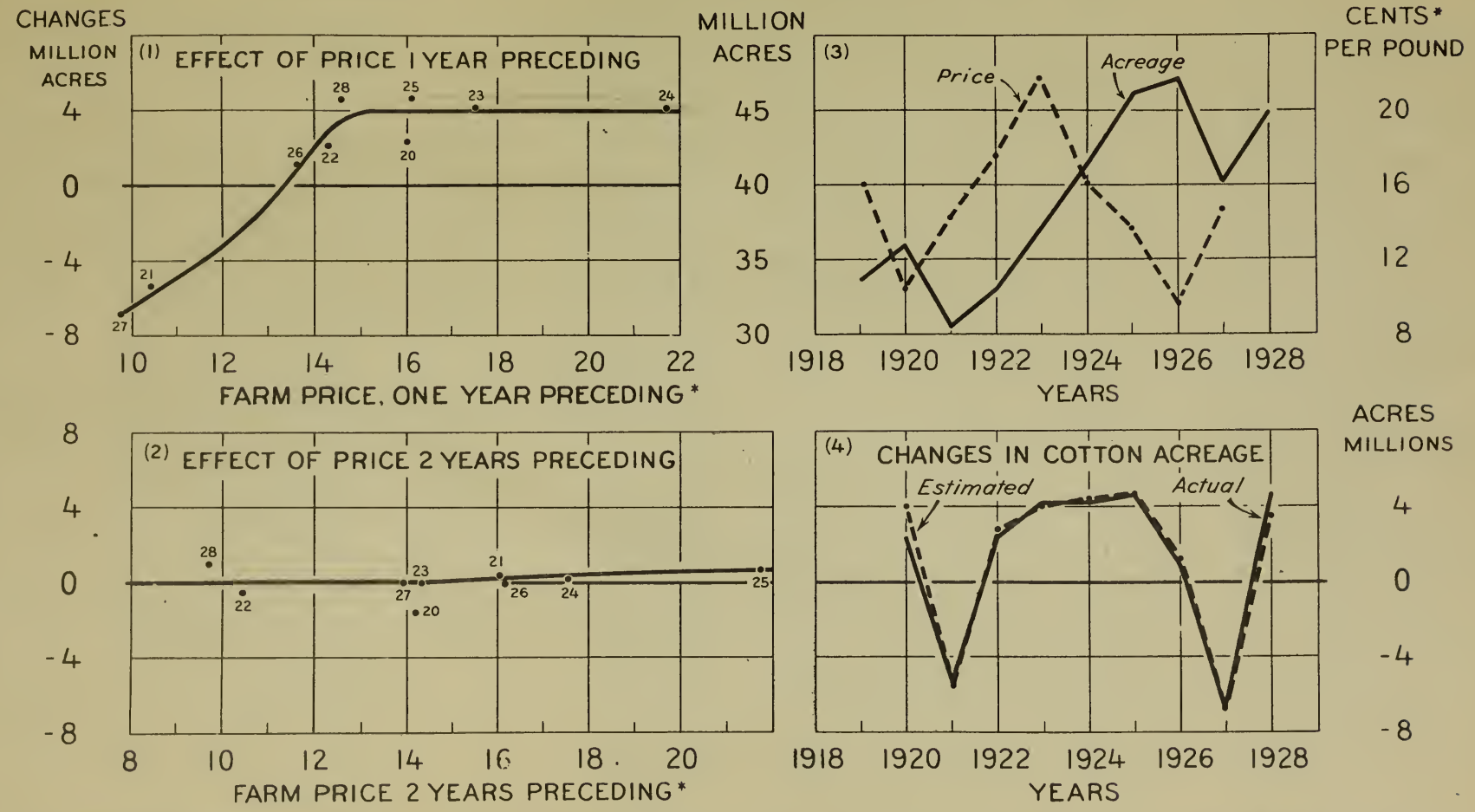

EFFECT OF SUPPLY AND OTHER FACTORS ON THE NEW YORK PRICE OF CALIFORNIA ORANGES
PRICE PER BOX IN 1926 DOLLARS
(1) EFFECT OF TOTAL PRODUCTION
TOTAL PRODUCTION - MILLIONS OF BOXES
INDEX OF PRODUCTION OF COMPETING FRUITS
(4) RELATION BETWEEN PRODUCTION OF ORANGES AND COMPETING FRUITS
PRICE DEVIATIONS FROM (1) DOLLARS
(2) EFFECT OF PRODUCTION OF OTHER FRUITS
FRUIT PRODUCTION, DEVIATIONS FROM (4)
PRICE DEVIATIONS FROM (2) DOLLARS
(5) TREND IN DEVIATIONS FROM (2)
YEARS
ACTUAL PRICE, PER CENT OF PRICE ESTIMATED IN 1926 DOLLARS
(3) EFFECT OF FOOD PRICES
INDEX OF FOOD PRICES, 1926 = 100
DOLLARS
(6) PRICES AT NEW YORK
Actual
Estimated
YEARS
U S DEPARTMENT OF AGRICULTURE
BUREAU OF AGRICULTURAL ECONOMICS

Figure 5

CHANGES IN U.S. COTTON ACREAGE, 1920-1928



U.S. DEPARTMENT OF AGRICULTURE

*FARM PRICE DIVIDED BY INDEX OF FARM PRICES, 1910-1914=100

BUREAU OF AGRICULTURAL ECONOMICS

Figure 6

and 1920, 1927, 1923. The deviations from this average relation between orange production and production of competing fruits may now be plotted against the price deviations from I. This is done in section 2. In section 2 the location of observations for 1923 and 1922 and for 1924, 1925, 1926, and 1927, indicates that the net effect of variations in competing production on the price deviations is a downward curve similar to the one shown. Study of the location of observations in relation to this curve or a similar curve in any other vertical position again reveals an upward trend in deviations which we may now plot in section 5 and pass through them a smooth upward sloping curve. As in the other cases, or examples, residuals from V should be plotted against each of the other curves to see if any changes are needed in their net shapes.

Thus we have in curves I, II and V (as finally modified, if necessary) practically a complete explanation of the price of oranges in terms of 1926 dollars. The final step in this analysis is to obtain estimates of orange prices which we may compare with the actual prices. Inasmuch as the actual prices were divided by the food index, we desire to express graphically the assumed relationship of the food price index to orange prices. As in the preceding example, this is obtained by plotting the index against ratios obtained by dividing the actual price by the sum of readings from curves I, II and V. These readings, it should be clear, may be taken as explaining the variations in orange prices due to all factors here dealt with other than those represented by changes in the food price level. The sum of the readings from I, II and V when multiplied by readings from this curve III give the desired price estimates which can be compared directly with the actual prices, as is shown in section 6.

The illustrations so far have dealt with relatively simple types of curves which describe the effect of one variable on another. In the last case the curves to be developed, namely, the effect of price on subsequent changes in acreage (case VI) are somewhat more complicated, but their essential characteristics are easily revealed.

CASE VI

Effect of price on acreage of cotton harvested in the United States

This example illustrates the application of a simple approach to curvilinear correlation in cases where the dependent variable is expressed in first differences or in percentage changes from one year to the next, such treatment being the best approach in analyses of acreage changes. In these analyses it is usually found that changes in one variable (acreage) from one year to the next respond to the price received by producers for the preceding and second preceding crop. Thus the changes in cotton acreage from one year to the next, and not the absolute acreage can be explained by prices, low prices in one season tending toward reduction in acreage and high prices toward expansion.

The method used in analyzing a case of this sort is shown in Figure 6. Section 3 contains the absolute acreages of cotton and the average price received by producers, the price used here being adjusted for changes in the general level of farm prices (1910-14 = 100). In section 4 are shown the changes in acreage from 1921 to 1928 from one year to the next. These acreage

changes are next shown in section 1, plotted against the price received during the year immediately preceding. Thus, in 1924 the reduction in cotton acreage of nearly 4.2 million acres is plotted against an average price received for the 1923 crop of 21.7 cents.

A tentative curve was then drawn in of a type which is characteristic of the effect of price on subsequent changes in acreage of such crops as potatoes, sweet potatoes, cotton, cabbage and wheat. This type of curve indicates that high prices in any given year result in a limited expansion of acreage, but higher prices do not produce any greater expansion. Reductions in acreage of the above-mentioned crops (except wheat) due to low prices are not as limited the first year as are increases, and lower prices bring still greater reductions. Residuals from that tentative curve were next plotted in section 2 against the price secured two years preceding the year of acreage changes and this indicated approximately a linear relationship. After studying the first distribution of residuals in section 2, adjustments were made in section I so as to obtain residuals for section 2 which would give a minimum of deviations from a curve in section 2. The prices two years preceding indicate a slight additional increase in acreage for very high prices, but no further decreases for low prices. As in the preceding cases readings from curves I and II give the estimates in section IV[8/].

In studying the relation of price to changes in acreage of other crops it will be found that the residuals from I (effect of price one year preceding) need other factors (such as prices three years preceding, prices of competing crops, weather, trends, etc.), in addition to the price two years preceding for their complete explanation, or reduction to a minimum. But these additional factors can be handled by methods already illustrated in the preceding cases.

CASE VII

Application of simplified methods to a general problem in multiple curvilinear correlation

The methods of graphic correlation used in the foregoing specific cases may be summarized by applying them to a general problem typical of the cases that are most often encountered in actual practices. For this purpose we take the data given in Table 3 for four variables X_1, X_2, X_3, X_4, and 30 sets of observations. The variations in X_1 are such that they correlate perfectly with X_2, X_3, and X_4. Our problem is to apply the simplified method of obtaining directly by inspection the net relationships between each of the three independent variables X_2, X_3, and X_4, and X_1 the dependent, without the use of the usual mathematical procedure. When this has been done we shall compare the results with those obtained by the mathematical procedure, and with the true relationships and shall find that the simple approach gives in much less time and labor practically the same net curves.

8/ For a more detailed analysis of changes in cotton acreage see Factors affecting cotton prices, U. S. Department of Agriculture Bulletin No. 50, by Bradford B. Smith.

The entire process is contained in the two accompanying Figures, 7 and 8, except the element of simple judgment which is required in studying or inspecting the independent variables before drawing the first approximations of the net regressions and no mathematical computations are involved other than the simple one of reading or measuring distances from curves and plotting such deviations in subsequent scatter diagrams.

In the following pages the steps of the simple procedure will be restated in the terms of the present problem of four variables. All the details will be given so that generalizations will be unnecessary. We may, however, note again that instead of establishing tentative net linear regressions by mathematical correlation we shall make use of scatter diagrams only and by inspection determine directly a tentative, but very close, approximation to the true net curvilinear regressions which will require only minor changes in the form of final approximations also to be made graphically.

The only computations involved are those required to obtain the index of correlation, but this is relatively simple, calling only for the sum of readings from the final curves (the usual $\bar{X}^1$) subtracting them from the actual values (X_1) computing standard deviations for the actual values of X_1 and for the final residuals ($\bar{X}^1 - X_1$) and substituting these in the formula for the index of correlation (p).

The steps now to be indicated in detail are:

1. Plotting three scatter diagrams, X_1 with X_2, X_1 with X_3 and X_1 with X_4 to determine by inspection if possible which of the three independent variables is the most important in the variations in X_1.

2. Determining by inspection a first approximation to the net relation between X_1 and X_2.

3. Determining by inspection which of the remaining two variables X_3 or X_4 is the more important in the X_1 variations not accounted for by X_2 and plotting against it (X_4) the residual variations from X_1 X_2.

4. Determining by inspection the first approximation to the net relation between X_4 and the residuals from X_1 X_2.

5. Plotting the residuals from the curve established in 4 against X_3 and determining the relation of X_3 to these final residual values of X_1.

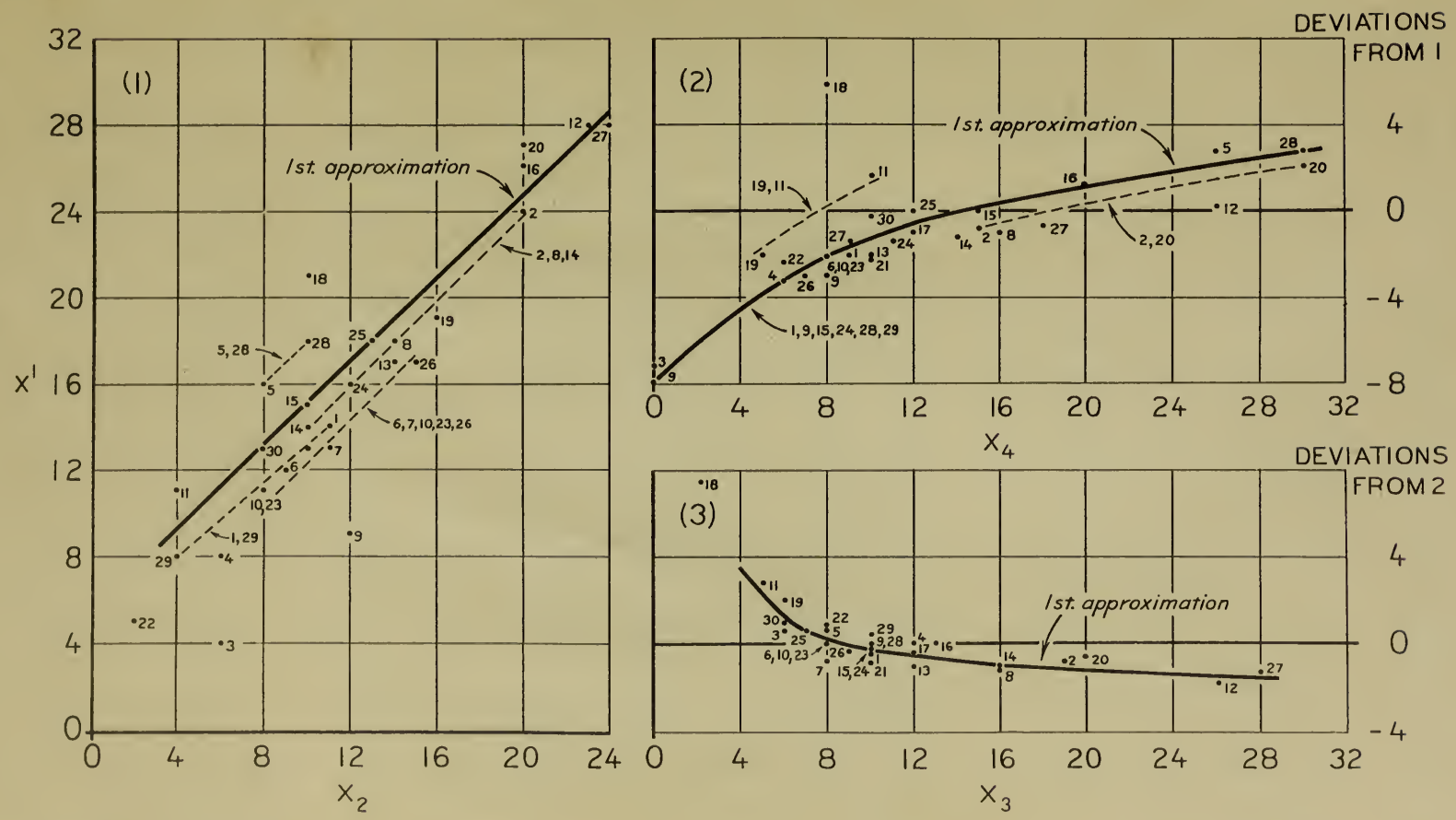
6. Plotting the residuals from the curve established in 5 as deviations from the other two first approximation curves and making second approximations, where necessary to reduce the residuals still further.

By plotting in the form of scatter diagrams X_1 and X_2, X_1 and X_3, X_1 and X_4 it became evident that the correlation between X_1 and X_2 is greater than between X_1 and either X_3 or X_4. We therefore select the X_1 X_2 scatter diagram and proceed to find the nature of the relation between X_1

and X_2. That diagram is shown in Figure 7. (As the other two scatter diagrams are not necessary hereafter they are not presented here). In Figure 9 we have also plotted the variables X_3 and X_4, having numbered them consecutively from 1 to 30, which we shall need to "inspect" as we proceed to determine the several net regressions. It should be observed that it is essential in our procedure that the identity of the individual observations be maintained.

In studying the scatter diagram X_1 X_2, we need to answer the question, Is the relation between X_1 and X_2 (with X_3 and X_4 constant), positive as indicated in the diagram or negative? Is it linear or curvilinear? To answer these questions, we make use of the fact that if the relation between X_1 and X_3 and X_1 and X_4 could be held constant simultaneously for two or more observations, the comparable observations in X_1 X_2 would lie along a line either linear or curvilinear which would indicate the true regression for X_1 X_2 for those two or more observations only. Now X_3 and X_4 in any two or more observations would bear a constant relation to X_1 under either of these two conditions, (1), if the X_3 values and the X_4 values were all equal or (2) if the X_3 values were equal and the X_4 values were equal. We therefore proceed to inspect the actual values of X_3 and X_4 for such combinations (see Figure 9) and note first that the observations numbered 6, 7, 10, 23 and 26 show equal values for both X_3 and X_4. We next find the comparable observations, 6, 7, 10, 23, and 26 in the X_1 X_2 scatter diagram and note that they appear to lie along a straight line, which is tentatively drawn in. Further inspection of X_3 and X_4 reveals that in observations 1 and 29 and 2, 8, 14, they have approximately the same values. We find and connect the comparable observations in X_1 X_2. We also note that in observations 5 and 28, X_3 has low but nearly equal values and X_4 has high but nearly equal values. As before we find and connect the 5th and 28th observations in X_1 and X_2. From the fact that the several lines so drawn and distributed through the diagram are nearly parallel, it is evident that the true regression of X_1 X_2 is a straight line of the slope indicated by the parallel lines and we proceed to draw a first approximation of that regression line through the body of the scatter diagram. (Had the true regression been curvilinear, a line connecting more than two observations for constant values of X_3 and X_4 would have revealed it.) This first approximation may now be taken as the tentative measure of the relation between X_2 and X_1 to be modified later if necessary and the vertical deviations from this regression may be assumed to be related to X_3 and X_4.

Our next step involves measuring or reading the differences between X_1 and the X_1 X_2 tentative regression, and plotting them against either X_4 or X_3. It is immaterial which of these independent factors are used first, but for convenience we may choose the one which appears to have the greatest influence on the X_1 residuals. Note that the greatest negative deviations from the X_1 X_2 regressions such as numbers 3 and 29 are associated with very small values of X_4 for those observations, and the greatest positive deviations 5, 20, and 28 are associated with very large values of X_4. These facts suggest that X_4 may be the dominant factor in determining the positive and negative residuals. They also suggest that the relation to be expected between X_4 and the residual values of X_1 is of a positive character. Incidentally this method of inspection also throws some light on the nature of the relation of X_3 on the residual values of X_1.



U. S. DEPARTMENT OF AGRICULTURE

BUREAU OF AGRICULTURAL ECONOMICS

FIGURE 7

For example, we note that number 18 among the observations in X_1 X_2 is well above the X_1 X_2 regression but instead of being associated with a high value for X_4 as in the other instances of large positive residuals it is associated with a low value of X_3. This suggests that the relation between X_3 and X_1 may be of a negative sort (at least for low values of X_3).

By plotting the vertical deviations from X_1 X_2 against X_4, as is done in Figure 8, Section 2, we obtain a scatter diagram in which we desire to discover the nature of the relation of X_4 to the residual values of X_1 (that is, to X_1 from which the effects of X_2 have already been removed). Inasmuch as these residual values in Section 2 are related to X_4 and X_3, we may proceed to find the relation of X_4 to the X_1 residuals by selecting those observations in which X_3 values are equal. In the observations numbered 1, 9, 15, 24, 28, 29, the X_3 values are equal. Connecting the comparable observations in Section 2 we obtain a curve of a positive slope, which appears to fit the scatter very well.

The adequacy of this curve may now be checked by selecting constant values of X_3 for large values of X_4 and also for low values. Consequently we note that in observations 2 and 20, X_3 has equal values. Connecting the two corresponding points in Section 2, we obtain a portion of a curve which is approximately parallel to the curve (for the high values of X_4) already drawn in through observations 1, 9, 15, 24, 28, 29. Similarly the X_3 values in 19 and 11 are practically equal, and a line connecting the comparable points in II are approximately parallel to the first curve (for the low values of X_4). If now we take the first curve (drawn through 1, 9, ---29) as the first approximation of the net relation of X_4 to the X_1 residuals, we note that many of the observations in Section 2 do not lie on that curve, presumably because of the influence of X_3. The influence of X_3 may now be observed by plotting the difference between the observations and the tentative curve in Section 2 against the comparable values for X_3. This step is shown in Section 3. The nature of the relation of X_3 to the residual values of X_1 is immediately evident. Instead of the positive gross relationship indicated by the scatter diagram of X_1 X_3 made at the beginning of the analysis, we now find a negative net regression, particularly pronounced for low values of X_3.

It is evident from the relatively narrow scatter of the observations in Section 3 about the first approximation curve drawn through them that by means of the three net regression curves developed so far we have accounted for nearly all of the variations in X_1. It remains now to see if some slight adjustments in these curves will reduce the scatter of the final residuals about the X_3 curve in Section 3 still more. At this point, if desired, the standard deviations from the X_3 curve in section 3 and the standard deviations of the original values of X_1 may be computed to determine the extent to which the three net curves account for the variations in X_1. Substituting these standard deviations in the formula for P, an index of correlation of .997 is indicated, the standard deviation of the final residuals being .46.

We may now complete the analysis by making a final test of the adequacy of the first approximation net regression curves. Here we follow the usual procedure of plotting the final residuals (as shown by the scatter around the curve in section 3) as deviations from the curve in section 1. This step is shown in Figure 8 to which have been transferred the first approximations from Figure 7. The scatter of the residuals about the X_1 X_2 curve indicates that no material adjustment in the shape or slope of that curve is necessary. The residuals are next plotted as deviations from the first approximation curve in Section 2. Here the scatter about the X_4 curve (in section 2), does indicate that a slight raising of the first approximation curve for the higher values of X_4 as well as for the very low ones, would reduce some of the residuals still more. This adjustment, drawn in by inspection, our second approximation for the X_4 curve is shown by the solid line in section 2. Had the scatter been wider about this curve it probably would have been desirable to follow the usual procedure of averaging or grouping, the deviations according to the values of X_4 in order to determine more exactly the shape of the second approximation curve. The scatter about this second approximation X_4 curve now indicates how much of the variations in X_1 can be accounted for by the three curves, (first approximation of X_1 X_2, second approximation X_1 X_4 and, first approximation of X_1 X_3)

The reduced residuals, that is, the deviations about the second X_4 curve are next plotted as deviations about the first approximation X_3 curve in Section 3 (Figure 8), to test the adequacy of that curve. This scatter indicates the desirability of lowering somewhat the first approximation X_3 curve. This adjusted curve now becomes the second approximation X_3 curve.

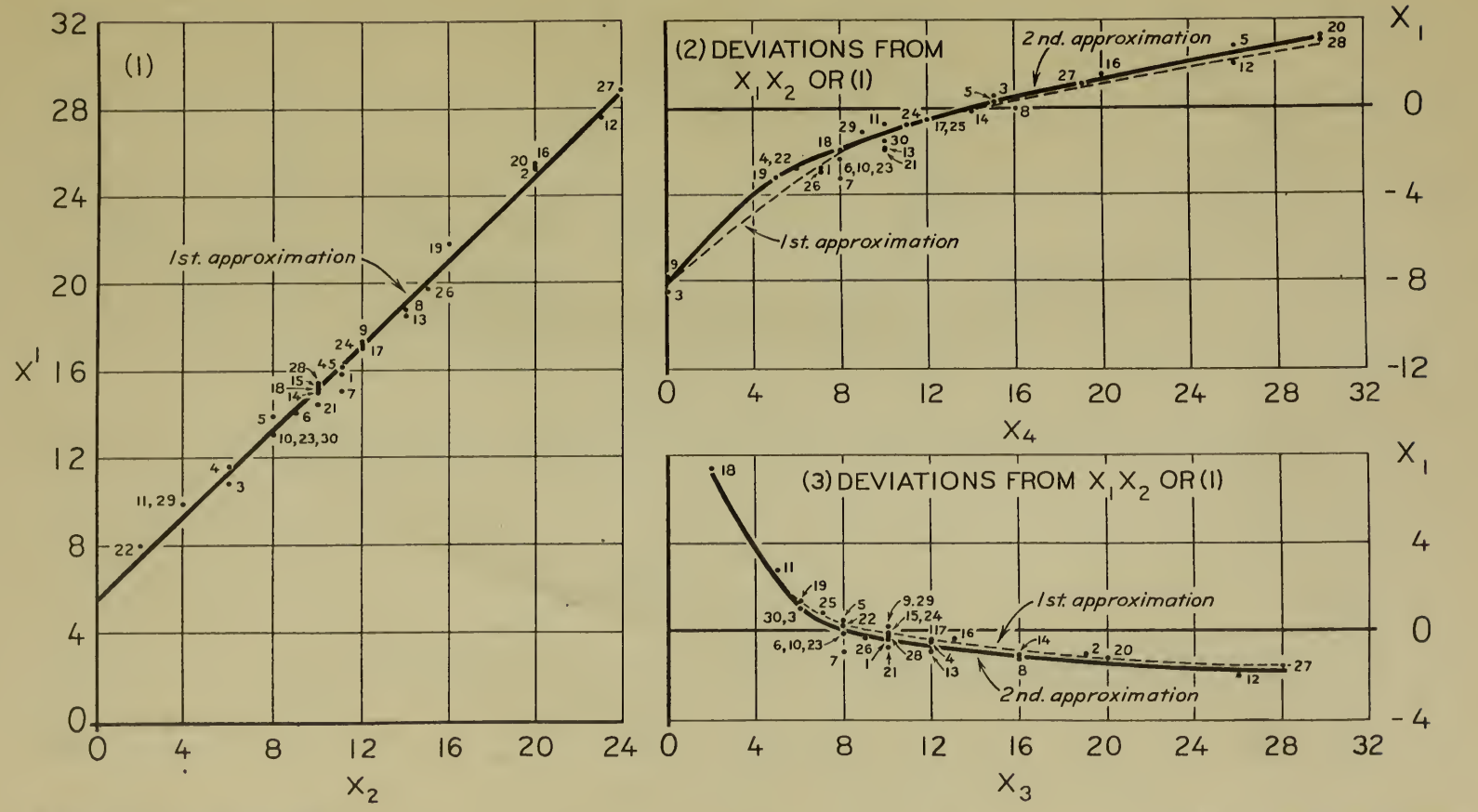
The extent to which these two adjustments have reduced the first set of residuals may now be seen either in the extent of the deviations about the X_3 second approximation X_3 curve or by computing values for X_1 from the three final net regressions. The readings from these curves, and the differences between them and the actual values of X_1 are given in table 3. The standard deviation of these differences, or final residuals is .411, with .46 obtained from the first and the index of correlation is .998 (compared with .46 and .997, respectively for the readings from the first approximations).

Comparison between the approximation curves and the true curves.

In order to determine whether the results obtained by the simplified approach to curvilinear correlation are reasonably accurate, we may compare them with the true curves and with the approximations that are obtainable by the usual method which involves the mathematical determination of linear net regressions and the reduction of residuals to a minimum by successive approximations, as described by Ezekiel.9/

To facilitate this comparison we used in this general illustration the data given by Ezekiel in his "Method of Handling Curvilinear Correlation" 1/ which are described by the formula $X_1 = X_2 + \frac{20}{X_3} + 2\sqrt{X_4} - 5$. The true net curves for X_1 X_2, X_1 X_3 and X_1 X_4, derived from this formula

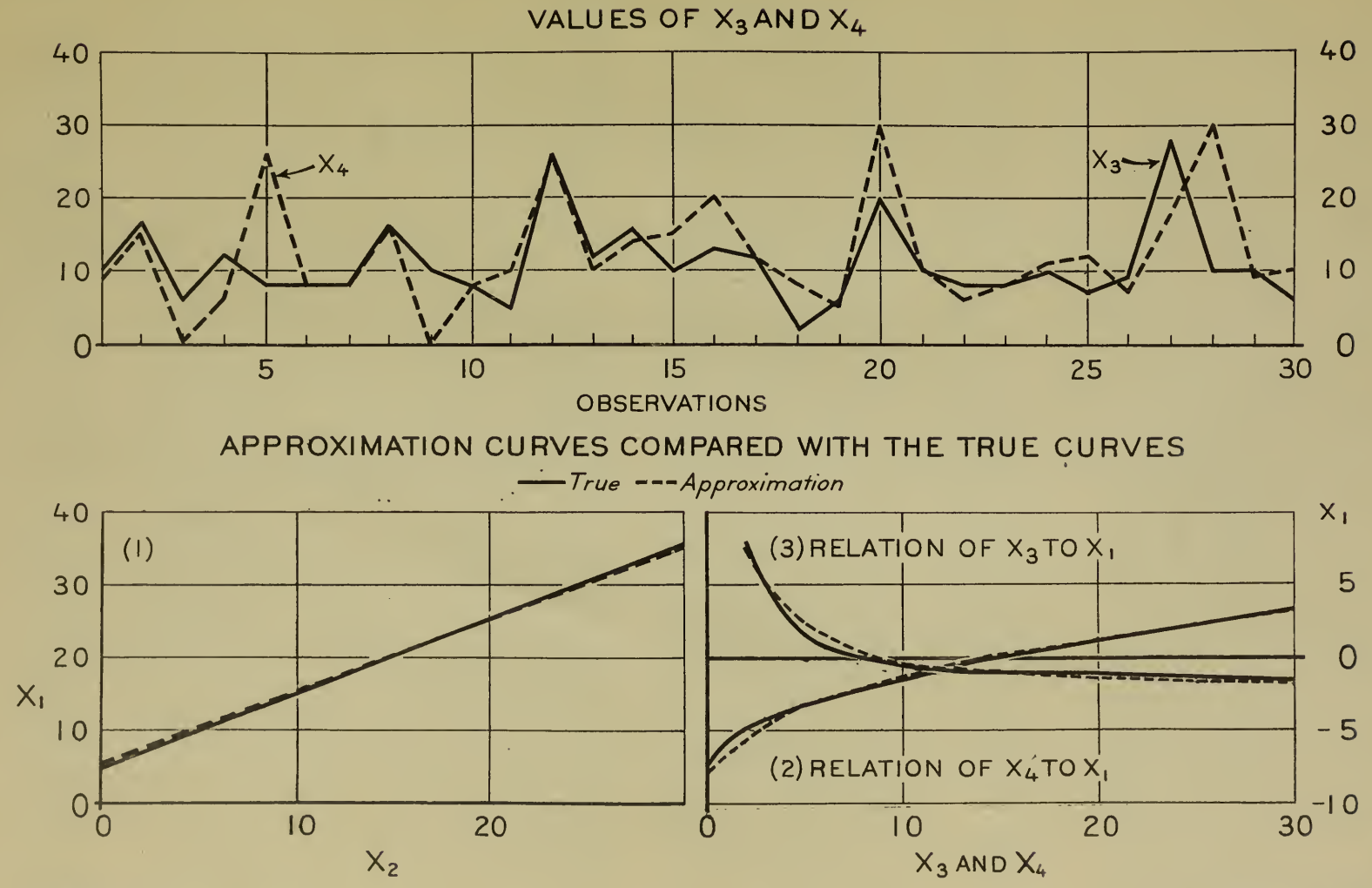
9/ See Journal American Statistical Association, December 1924.



U.S. DEPARTMENT OF AGRICULTURE

BUREAU OF AGRICULTURAL ECONOMICS

FIGURE 8



U. S. DEPARTMENT OF AGRICULTURE

FIGURE 9

BUREAU OF AGRICULTURAL ECONOMICS

are indicated in Figure 9 and are compared with the curves obtained by the simplified method of correlation. The approximation curve in section 1 has only a slightly different slope from that of the true curve. The approximation curve in section 2 does not rise as much for values of X_4 between 0 and 5, as does the true curve, but this is due to the fact that the data used in this problem had no values for X_4 between 0 and 5. The approximation in section 3 also differs only slightly from the true curve.

These close agreements between the true curves and the approximations may be compared with the results obtained by the Ezekiel Method, by referring to page 446 of the December, 1924 Journal of the American Statistical Association. It will be observed that the approximation curves there derived show in general practically the same agreement with the true curves. For the curves X_3 X_1 and X_4 X_1 the agreement is somewhat closer as developed here (in Figure 9).

A correlation index of .994 was there obtained after three successive approximations which may be compared with the correlation index of .998 after only one adjustment as indicated above.

The results obtained by this simple approach in the first six illustrations are also practically the same as those obtainable through the usual procedure. The data in Case II, for example, were correlated by the usual method, which required a series of four approximations to obtain a final correlation of .993. The simple approach gave in much less time the same correlation, .995, and practically the same net relationships between price and consumption, between business activity and consumption, and the trend in residuals as were obtained by the mathematical correlation.

Both methods of curvilinear correlation depend to some extent on judgment. In the usual method, judgment comes into play in converting linear regressions into curvilinear ones. In the simple approach, judgment is brought into play in the process of determining first approximations to net curvilinear regressions directly by inspection. In both cases there is some freedom in shaping the curves. Where two independent variables are highly correlated, and one of the net regressions is given an inordinate slope, it will be compensated by a corresponding change in the other curve. For example, had we drawn in Figure III, section 5, a curve instead of a straight line, there would have appeared a compensating difference in the slope of the trend line in section 6 without any effect on the final correlation. As a matter of fact, in this particular instance, the simpler approach reveals immediately that the assumption that the true relationship is a straight line will agree with the observations, while the usual approach without a similar process of inspection would lead one to assume an illogical curvilinear function.

The question may be raised whether the simple approach can be conveniently used in problems involving more observations and more variables than those in the illustration.

Inasmuch as the facility of this method depends on detecting approximate net regressions by inspection instead of by mathematical computation of linear regression and successive approximation, too many

observations in a scatter diagram are likely to make it difficult to find the true relationship. But this limitation can be overcome by splitting the problem into two or more sections and treating each separately. Something of this sort was indicated in example IV where the average net effect of supply on the price of apples was determined by establishing tentative curves for the pre-war and post-war years. Such treatment of a long time series of observations is in fact likely to give much truer and more reliable relationship, particularly if the same or similar curves are found to hold good in each of two or more periods.

For problems of 30 observations and variables such as that treated by Ezekiel in illustrating his method of curvilinear correlation[4] the simple approach is, as we have seen, very satisfactory. That set of data when treated by methods described here gave the same net curves and the same high correlation, but the time involved was only about one-fourth as much.

Another question concerns the significance or reliability of the results obtained by this method for such short time analyses as have been presented here particularly in Cases I to VI. We have two tests that may be applied.

One is to repeat the analyses for another period or for a longer one to determine whether similar results will be obtained. This, however, assumes that economic relationships remain unchanged from one period to the next, which is not necessarily so. When an analysis for an earlier period corroborates the result of a more recent one, it lends greater confidence in the latter; but if it does not agree, it does not invalidate the latter, for different sets of forces may be at work in one period than in another. Thus, in the acreage analysis referred to on page 10 the world war and the boll weevil were factors in the earlier part of the period, but not in the more recent years. However, the net effects of price on acreage in case VI are of the same type as those derived by the more detailed study based on the formal approach.

A second test is the practical one of applying the results obtained for a short period to the year or years immediately preceding or following. In each of the cases presented and a number of others that might have been presented, very satisfactory results were obtained when the relationships established for the period ending with 1927 were applied to 1928. This, however, is like the preceding test in that if the relationships established for a given period hold also for a year outside that period, we may have greater confidence in the established relationships, but they are not necessarily invalidated if they do not apply with equal accuracy to outside years.

Finally, for those who are accustomed to thinking of goodness of fit and reliability of results in terms of correlation coefficients, correlation indexes, and standard errors, it may be of interest to point out how nearly we have accounted for all the variations in the dependent variables in each of the foregoing problems.

For this purpose we may make use of the correction that must be applied to correlation coefficients, taking into account the number of

variables or constants determined in the regression equation, and the number of variables the need and significance of which has already been described in the Journal 10/ of the American Statistical Association. The correction indicated to what extent the observed correlation coefficients in a sample may overstate the true correlat on existing between the same variables in the universe from which they were selected. When applied to a correlation in time series, correction for the number of variables and observations does not have the usual significance, for a time series is not a sample. However, even though correlation coefficients in time series lack the significance that they have in problems where different samples may be drawn, some readers may be interested in the correlation indexes determined for the six foregoing cases as well as the seventh.

In the following tabulation are given first the original multiple correlation indexes and the standard deviations of the residuals, and in the last two columns, the indexes and standard errors, after correcting for the number of observations and the maximum number of variables and constants that may be assumed to be represented in the net curves.11/ In these cases of very high correlations the corrections for the number of variables and constants are not material, and, in so far as this criterion is concerned, the corrections do not impair the validity of our results.

Case	P	$O^{-}z$	Number of variables and constants	Number of observations	$\bar{P}$	Se
I.....	.997	2.67	5	8	992	3.38
II.....	.995	1.0	4	10	991	1.3
III.....	.995	1.0	5	10	989	1.4
IV.....	.998	2.37	5	18	997	2.79
V..1/....	.997	.07	7	8	978	.20
VI.....	.986	.69	5	9	968	1.04
VII.....	.998	.41	6	30	997	.46

1/ Based on the estimates in terms of prices adjusted for changes in the food price level.

10/ Proceedings of the American Statistical Association, December, 1928. Paper by M. J. B. Ezekiel, on Application of Theory of Error to Multiple and curvilinear correlations.

11/ The corrections are made by the use of these two formulae:

$$\text{Corrected } P = \bar{P}^2 = 1 - \frac{1 - p^2}{1 - \frac{m}{n}} \quad \text{and}$$

$$\text{Corrected } O^{-}z = Se = n\,\frac{O^{-}z}{n-m}$$

The formula for $\bar{P}$ is that developed by B. B. Smith and applied to curvilinear correlation by M. J. B. Ezekiel. The formula for Se is R. A. Fisher's formula (Statistical Methods for Research Workers, p. 135) restated by M. J. B. Ezekiel.

Table 1 - Data used in cases I - IV

Year	Case I: Production 1/	Case I: Price per bush. received by producers	Case I: May price per cwt. of old potatoes Chicago	Cases II and III: Index of cotton consumption 2/	Cases II and III: Price per lb. of cotton	Cases II and III: Index of production of mfrs. 3/
	Million bushels	Dollars	Dollars	Per cent	Cents	Per cent
1919	--	--	--	96	22.0	84
1920	--	--	--	95	24.7	86
1921	21.2	1.12	.87	88	14.5	66
1922	24.1	1.34	1.70	99	18.2	87
1923	18.7	1.67	1.13	106	25.3	101
1924	29.4	.99	1.50	89	30.6	94
1925	20.4	1.41	1.13	105	24.6	105
1926	23.7	1.72	3.23	109	19.7	108
1927	29.6	1.55	3.51	122	15.3	106
1928	37.4	.65	1.43	107	20.4	111

1/ Potato production 10 early states.
2/ 1923-25 = 100, Federal Reserve Board.
3/ At New Orleans, crop year ending in the indicated calendar year.

Year	Case IV: Price per bush. received by producers 4/	Case IV: Total production of apples	Case IV: Index of food prices 5/	Case IV: Ratio of actual prices to prices read from curves I and II
	Dollars	Million bushels	Per cent	Per cent
1910	102.6	142	61.8	61.8
1911	91.1	214	65.7	71.7
1912	74.8	235	65.0	63.9
1913	106.1	145	63.9	64.7
1914	71.7	253	66.5	65.8
1915	79.4	230	67.8	69.6
1916	104.2	194	89.3	89.8
1917	125.9	167	112.6	112.4
1918	154.6	170	126.0	125.7
1919	208.9	142	137.5	135.6
1920	144.2	224	111.1	110.9
1921	197.4	99	86.8	87.3
1922	130.4	203	91.6	91.2
1923	125.0	203	90.7	91.2
1924	138.7	172	95.8	96.3
1925	137.4	172	101.5	100.3
1926	99.1	247	97.4	102.1
1927	156.4	123	98.6	97.1

4/ Straight average July-May 5/ 1926 = 100 Bureau Labor Statistics average for July-June

Table 2 - Data used in Cases V and VI

Year	Case V: Changes in cotton acreage	Case V: Price per lb. received by producers 1/	Case VI: N. Y. price per box of Calif. oranges: Nov. - Oct.	Case VI: Price in terms of 1926 Dollars 2/	Case VI: U.S. production of oranges	Case VI: Index of production of competing fruits 3/
	Million acres	Cents	Dollars	Dollars	Million boxes	Per Cent
1918	--	14.2	--	--	--	--
1919	--	16.0	--	--	--	--
1920	+ 2.3	10.4	5.75	5.18	29.9	94.1
1921	- 5.4	14.3	7.19	8.28	20.1	83.6
1922	+ 2.4	17.5	5.29	5.78	29.9	107.9
1923	+ 4.1	21.7	5.39	5.94	34.2	106.7
1924	+ 4.2	16.1	7.01	7.32	28.0	97.5
1925	+ 4.7	13.7	6.15	6.06	31.0	113.3
1926	+ 1.0	9.7	5.58	5.73	35.7	125.4
1927	- 6.9	14.6	6.60	6.70	33.5	103.2
1928	+ 4.7					

1/ Weighted average farm price August-July divided by July-June index of farm prices 1910-14 = 100.

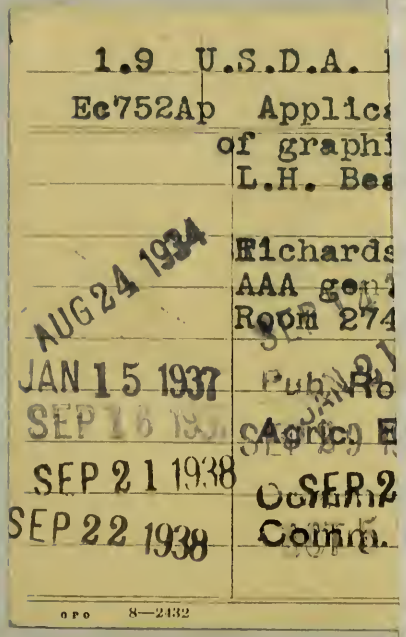
2/ New York price divided by index of food prices, see column 3 under case IV.

3/ 1919-1927 = 100, index includes production of apples of one year and of peaches, pears, strawberries and grapes of the next weighted by average prices received during 1919-1927.

Table 3 - Data used in Case VII

Item number	Raw data				Readings from curves					Item number
	X_2	X_3	X_4	X_1	I	II	III	Sum ($\bar{X}_1$)	e	
1	11	10	9	14	16.2 -	2.5	- .3	13.4	+ 0.6	1
2	20	19	15	24	24.9 +	0.2	- 1.3	23.8	+ 0.2	2
3	6	6	0	4	11.3 -	8.1	+ 1.2	4.4	- 0.4	3
4	6	12	6	8	11.3 -	2.8	- 0.6	7.9	+ 0.1	4
5	8	8	26	16	13.3 +	2.5	+ 0.2	16.0	0	5
6	9	8	8	12	14.2 -	2.1	+ 0.2	12.3	- 0.3	6
7	11	8	8	13	16.1 -	2.1	+ 0.2	14.2	- 1.2	7
8	14	16	16	18	19.0 +	0.8	- 1.1	18.7	- 0.7	8
9	12	10	0	9	17.0 -	8.1	- 0.3	8.6	+ 0.4	9
10	8	8	8	11	13.3 -	2.1	+ 0.1	11.3	- 0.3	10
11	4	5	10	11	9.4 -	1.3	+ 2.4	10.5	+ 0.5	11
12	23	26	26	28	27.7 +	2.5	- 1.6	28.6	- 0.6	12
13	14	12	10	17	19.0 -	1.3	- 0.6	17.1	- 0.1	13
14	10	16	14	14	15.2	0	- 1.1	14.1	- 0.1	14
15	10	10	15	15	15.2 +	.2	- 0.4	15.0	0	15
16	20	13	20	26	24.9 +	1.3	- 0.7	25.5	+ 0.5	16
17	12	12	12	16	17.0 -	.7	- 0.6	15.7	+ 0.3	17
18	10	2	8	21	15.2 -	2.0	+ 7.4	20.6	+ 0.4	18
19	16	6	5	19	21.0 -	3.2	+ 1.2	19.0	0	19
20	20	20	30	27	24.9 +	3.1	- 1.4	26.6	+ 0.4	20
21	10	10	10	13	15.2 -	1.3	- 0.3	13.6	- 0.6	21
22	2	8	6	5	7.5 -	2.8	+ 0.1	4.8	+ 0.2	22
23	8	8	8	11	13.3 -	2.1	+ 0.1	11.3	- 0.3	23
24	12	10	11	16	17.0 -	.9	- 0.4	15.7	+ 0.3	24
25	13	7	12	18	18.0 -	.7	+ 0.6	17.9	+ 0.1	25
26	15	9	7	17	20.0 -	2.5	- 0.1	17.4	- 0.4	26
27	24	28	18	28	28.6 +	1.1	- 1.8	27.9	+ 0.1	27
28	10	10	30	18	15.2 +	3.1	- 0.3	18.0	0	28
29	4	10	9	8	9.4 -	1.7	- 0.3	7.4	+ 0.6	29
30	8	6	10	13	13.3 -	1.3	+ 1.2	13.2	- 0.2	30
$\bar{0}$			6.28						.411	

****



UNITED STATES DEPARTMENT OF AGRICULTURE
Bureau of Agricultural Economics

APPLICATIONS OF A SIMPLIFIED METHOD OF GRAPHIC CURVILINEAR CORRELATION

By

L. H. Bean, Senior Agricultural Economist
Division of Statistical and Historical Research

A Preliminary Report

Part II

The Method Applied to Changes in
Acreages, Yields and Livestock Numbers

Washington, D. C.
September, 1929

A SIMPLIFIED METHOD OF GRAPHIC CURVILINEAR CORRELATION APPLIED TO CHANGES IN ACREAGES, YIELDS, AND LIVESTOCK NUMBERS

By L. H. Bean, Senior Agricultural Economist, Division of Statistical and Historical Research, Bureau of Agricultural Economics

The simplified method of correlation already described in considerable detail elsewhere 1/ may be further illustrated by applying it to three additional problems dealing with actual changes in acreage, livestock numbers and yields. The cases selected for illustration deal with changes (a) in the United States acreage of cabbage, (b) in the total number of hogs on farms and (c) in the yield per acre of wheat in an eastern State. It will be noted that the first two of the present illustrations (VIII and IX) are similar to Case VI already described, in that the dependent variables (acreage and hog numbers) are expressed as first differences or absolute increases or decreases from the preceding year's figure. The final illustration (X) is like the general problem described under Case VII.

Case VIII. The Relation of Price to Changes in the United States Acreage of Cabbage

In this problem it is desired to correlate the price of cabbage received by producers with subsequent changes in acreage and to develop two curves, one representing the relation of the price received for the crop in the first year preceding the acreage change and another, the relation of the price received two years earlier. The prices used here have been adjusted for changes in the general level of farm prices.

Our first step is to plot in a scatter diagram the price one year preceding against changes in acreage (See section 1, Figure 10). The second step is to obtain for that scatter diagram an approximation to the relation of price one year preceding to acreage changes, exclusive of the influence of the price two years preceding. As an aid in making that approximation, we examine the variations in the price two years preceding for equal or approximately equal values of that variable and note (in Section 3-Figure 10) that the prices in 1920, 1924 and 1925 are approximately the same. Now if the price two years preceding has any influence on acreage, then these three similar prices should have about the same influences on the 1922, 1926, and 1927 acreage changes. In other words, their effect in these three years may be considered tentatively, as constant. Consequently, we may draw a line or curve through the 1922, 1926 and 1927 observations in Section 1, thus obtaining for that section of the diagram a partial indication of the nature of the net curve we are seeking. We note next that the prices of 1921 and 1923 are relatively high, the 1921 price being higher than the 1923 price. Inasmuch as the 1921 price is greater than the 1923 price, it is to be expected that its influence on the 1923 acreage may be greater than that of the 1923 price on the 1925 acreage. By connecting the 1923 and 1925 observations and allowing the curve to remain below the 1923 point because of the greater influence just referred to, we obtain another indication of the nature of the

---

1/ See Mimeographed Report, Part 1 on Applications of a Simplified Method of Graphic Curvilinear Correlation.

net curve or regression for another portion of the scatter diagram. With these two lines drawn in, it now becomes obvious that the 1924 acreage is below that indicated by the line passing through 1922-26-27, because the 1922 price was low.

The two segments drawn in so far may be taken to represent the relation between price one year preceding and acreage with prices two years preceding held constant respectively at the low prices of 1920, 1924, and 1926, and at the high prices of 1921 and 1923. They indicate that the slope of the curve for Section 1 rises sharply when prices one year preceding range between $12.00 and $16.00, and that for higher prices the curve slopes upward very moderately. Using these two segments as guides, we may draw a continuous free hand curve, as the first approximation to the net relation of price one year preceding to acreage changes. The solid curve shown in Section 1 is practically the first approximation.

Now if the first approximation curve in Section 1 represents the acreage changes that may be attributed to, or associated with the price one year preceding, the amounts of acreage change above or below that curve for the years shown may be assumed to be due to the influence of the second factor under consideration, namely, the price two years preceding. We therefore proceed to relate the price two years preceding to those portions of acreage changes not already attributed to the other price. This is most conveniently done by measuring off or reading directly the differences between the observations and the curve in Section 1, and plotting those differences against price two years preceding in Section 2 of Figure 10. All of the observations in Section 2 are found to lie along a fairly well defined curve (except 1921) and a free hand curve drawn through them may be taken as the first approximation to the relation between prices two years preceding and acreage changes.

Inasmuch as the observations in Section 2 show so little scatter about the first approximation curve, both of the curves, in Section 1 and in Section 2, may be taken as final. In cases where the scatter is wide, it is necessary to test the validity of the first approximations. This can be done by measuring the amounts that the observations in Section 2 are above or below the curve first approximated in 2 and plotting these deviations above or below the curve in 1. If these deviations, transferred from Section 2 to the first approximation in 1, group themselves at any point consistently above or below the first approximation curve in 1, the curve may be altered so as to reduce the deviations at that point. This gives a second approximation for Section 1. Deviations from this second approximation are then plotted around the curve in Section 2 and the curve there adjusted if the new residuals suggest it. This gives a second approximation curve for Section 2. If necessary this process is repeated until the residuals are reduced to a minimum.

Section 4 of Figure 10 shows the usual comparison between the actual acreage changes and those estimated from the two price-acreage curves.

Case IX. The Relation of Price to Changes in the Number of Hogs on Farms in the United States on January 1

This problem is similar to the preceding one in that the dependent variable, changes in the number of hogs on farms, is here taken as absolute first differences or changes from the numbers on farms on the preceding January 1, and in that the independent variables are two price factors, one being the corn-hog ratio for the first 12 months period preceding January 1 and the other, the corn-hog ratio for the second preceding 12-month period. The method of determining the curves for each of these price influences is similar to that shown for cabbage acreage in Case VIII. There is, however, one important difference, namely, that for the period under consideration, 1920-1929, there appears to have been a downward trend in the relation between the corn-hog ratio, and the number of hogs on farms. In presenting this problem, therefore, we shall refer only to this additional factor and indicate how its presence may be detected and its influence held constant in determining the relation of the other factors to the dependent variable.

In Section 1, Figure II, is shown the final approximation of the relation of the corn-hog ratio in the first preceding year on changes in hog numbers; in Section 2, the final approximation for the corn-hog ratio the second 12-month period preceding; and in Section 3, the trend in the relationship, or stated differently, the trend in changes in hog numbers, not attributed to or associated with the two corn-hog price series. We need to note only Sections 2 and 3. Having drawn the curve in Section 1 by a procedure similar to that already described in the preceding illustration and then having plotted in Section 2 the residuals from the curve in Section 1, the problem is to draw the first approximation curve for the effect of the corn-hog ratio two years preceding. Here it is found that the observations do not fall along a well defined curve and drawing the first approximation curve is not as simple as it was in Case VIII (Cabbage acreage changes and price two years preceding.) An inspection of the observations reveals first that the relationship is probably positive, that is, that the curve rises with higher corn-hog ratios, as in Section 1. It is next observed that any upward sloping line that may be drawn through the observations would leave those for the earlier years in the series above the line and those for the later years, below the line, indicating the presence of a downward trend that may be associated with time. Thus, the problem becomes one of three independent variables and time, the third variable, must be taken into account and its influence held constant in determining the nature of the relation between the corn-hog ratios and changes in hog numbers.

The method of holding time constant, in determining the first approximation in Section 2, is indicated by the dashed lines. The process is simply to connect the observations in chronological sequence, bearing in mind that if the trend factor is continuously downward, the connecting lines should not cross, but should fall in descending order (or ascending order where the trend is upward). When the observations in Section 2 are so connected the general nature of the relation of the second corn-hog ratio to hog numbers is sufficiently obvious, and a first approximation may be made which is not materially different from the final one (shown in the solid line). The curve shown here

has been arbitrarily placed so as to show most of the downward trend in residuals prior to 1927. It could of course have been placed higher, but the effect of that would have been merely to lower the curve in Section 3 in relation to the zero line in Section 3.

The next step in this type of problem, given a trend factor, is to plot in Section 3 the deviations from the first approximation in Section 2, and to pass through them a line of best fit. To test, finally, the goodness of fit of the three curves so developed, residuals from the trend line should be plotted as deviations from the other two curves in the usual manner.

For further applications of this method and a discussion of Cases VIII and IX the reader is referred to the Journal of Farm Economics, July 1929, "The Farmers' Response to Price" by the author.

Case X. The Relation of Three Weather Factors to Wheat Yields in State "X"

In this final illustration our object is to apply the simplified correlation method to a yield problem by correlating three weather factors, rainfall, snow cover and temperature with the yield of wheat in a certain State for a selected period of ten years. The purpose of this illustration is not so much to present the nature of the relation of each of these factors to yield, but rather to indicate how the simplified method may be applied to complicated yield problems, the analyses of which ordinarily consume a great deal of time. The weather factors used are (1) rainfall during February, March and April, (2) a measure of snow cover (the number of days of one inch or more of snow on the ground) and (3) average temperature in March and April. The years have been numbered 1 to 10 inclusive. 1/

The procedure followed in this illustration is practically identical with that described under Case VII in part I of this report, except for a slight modification in the device used to find sets of observations in which the influence of two factors appear to be approximately equal in order to obtain the first approximation of the influence of the third factor on yield.

The first step, shown in Section 1 of Figure 12, is to plot yield against one of the independent variables, (rainfall) and then to study the variations in the other two variables so as to obtain a first approximation curve for Section 1. Instead of plotting the two dependent factors consecutively, as was done in Figure 9 (cases VII), we make use of a scatter diagram (See Section 2 of Figure 12) with temperature plotted against snow cover. Inspecting this scatter diagram for two or more observations in which these two factors may be assumed to have approximately equal values, we note that (a) observations 8 and 9 have relatively low temperature and low snow cover

---

1/ The data for observations 1-9 inclusive used in this illustration were supplied by Mr. S. R. Newell of the Division of Crop and Livestock Estimates, who has successfully used these factors in forecasting the yields of the past two seasons. Data for observation number 10 are based on preliminary curves.

values, (b) observations 4 and 5 have relatively low temperature, but greater snow cover values, and (c) observations 3 and 6 have average or better than average temperatures, and still greater amounts of snow cover. The sets of observations in Section 1 comparable to these may now be inspected to obtain suggestions of the nature of the influence of rainfall. It should be observed that none of the sets of observations in Section 2 contained equal values, (for instance, 8 has higher temperature and lower snow cover than 9). Consequently the dotted lines in Section 1 do not connect the two observations in each set, but they nevertheless suggest the slope and shape of the first approximation curve.

The next step is shown in Section 3 of Figure 12, where deviations from the curve in Section 1 are plotted against snow cover. The shape of the first approximation curve is here revealed by connecting the observations in the order of the values of X_3, the factor that here needs to be held constant. Note that observations 7, 5, 9, 1, 4 and 8 are connected in sequence in the order given, because the corresponding values of X_3 are 44.0 45.5, 45.9, 47.0, 47.5, 47.7. The other observations are also connected in sequence, 3, 6, 10, 2, for the corresponding values of X_3 are 49, 50, 54, 56. Each set of dashed lines suggest the first approximation free hand curve shown in Section 3. Deviations from the free hand curve in Section 3 are then plotted in Section 4 against temperature, and a first approximation curve drawn through them.

To test the validity of the first approximation curves in Sections 1, 3, and 4, it is necessary to transfer the deviations about the curve in 4 to each of the other curves. This may be accomplished by plotting the three preliminary curves in Sections 5, 6, and 7 and then plotting deviations from Section 4 as deviations about the first approximation for X_1 X_2 in Section 5. This process suggests a somewhat steeper slope for the curve X_1 X_2. Consequently a second approximation curve is drawn. But inasmuch as the second curve still shows deviations to be accounted for, these need to be transferred to the X_1 X_4 first approximation in Section 6. Here too, the deviations suggest a slight change in the preliminary curve, namely, raising it for low values of X_4 and lowering it for high values, as shown by the second approximation. Finally, the deviations about the second approximation in Section 6 are transferred to the first approximation in Section 7, X_1 X_3 and the latter modified slightly to reduce the residuals numbered 1 and 3.

By this simple process three net curves, X_1 X_2, X_1 X_4 and X_1 X_3 are developed which account for practically all of the variations in yields for the 10-year period under examination, except about 1 bushel in the year marked "3".

If from this point on it is desired to compute correlation and determination coefficients, the usual procedure may be followed by treating the deviations from the second approximation curve X_1 X_3 in Section 7 as final residuals from which to compute the standard deviation required for the index of correlation formula.

Date used in Cases VIII and IX

Year	Case VIII: Average price per ton of cabbage received by producers 1/	Case VIII: Yearly changes in United States cabbage acreage	Case IX: Corn-hog ratio 2/	Case IX: Changes in number of hogs on farms January 1
	Dollars	1,000 acres	Bushel	Millions
1919.....	19.58	--	10.3	--
1920.....	16.26	+ 27.6	9.8	- 3.84
1921.....	28.52	- 19.2	14.0	- 1.36
1922.....	12.94	+ 29.2	14.4	+ .96
1923.....	23.28	- 28.9	9.0	+ 9.48
1924.....	16.04	+ 14.2	8.2	- 2.68
1925.....	17.02	+ .9	11.3	- 10.79
1926.....	19.03	+ 9.3	16.9	- 3.42
1927.....	15.97	+ 14.5	12.7	+ 2.64
1928.....	24.43	- 7.0	9.9	+ 5.63
1929.....	--	--	--	- 5.46

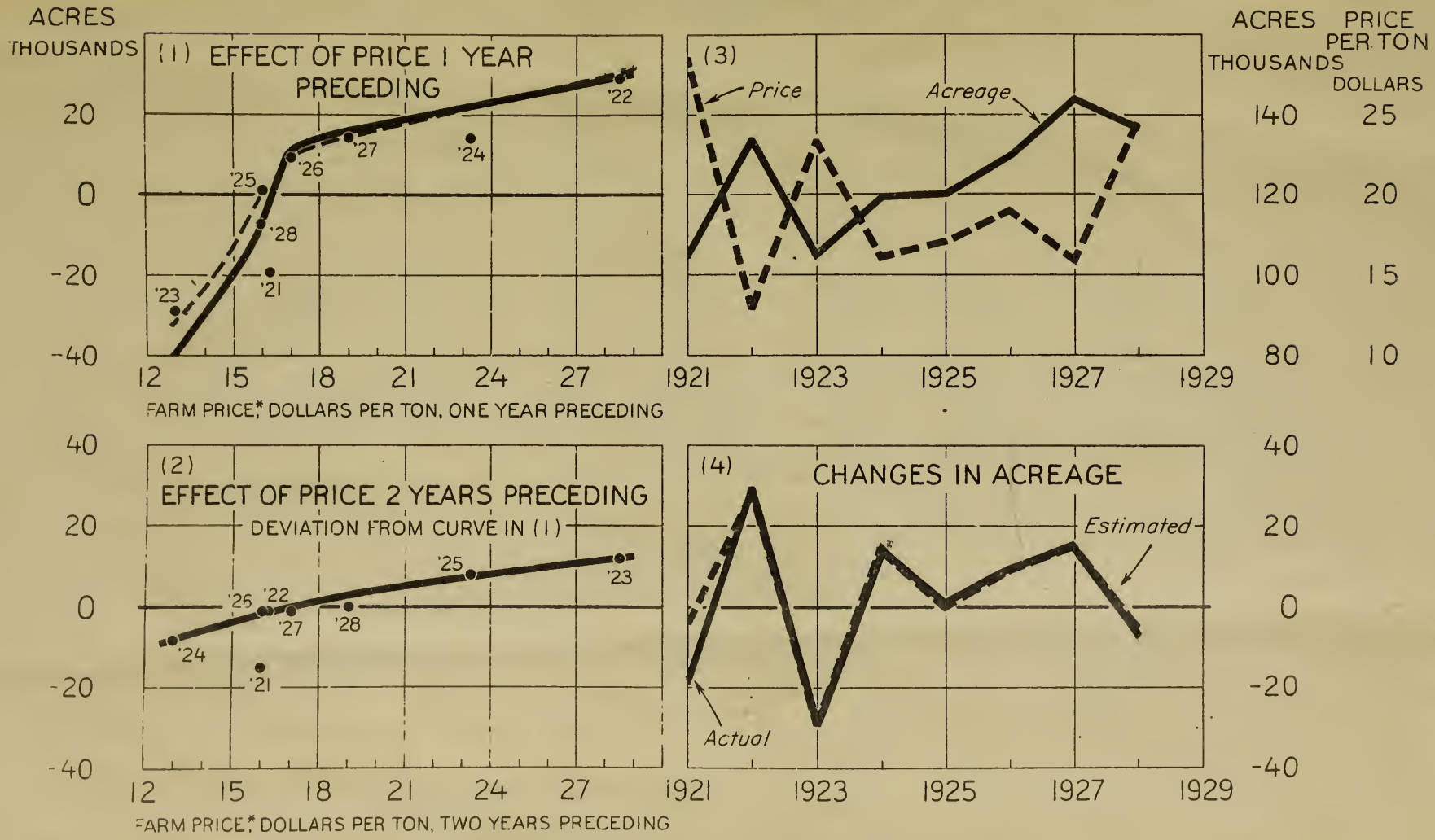
1/ Adjusted for changes in crop year index of farm prices, 1927-28 = 100.
2/ Farm price of hogs per hundredweight divided by farm price of corn per bushel, calendar year average.

Data used in Case X

Year	Rainfall 1/	Temperature 2/	Index of snow cover 3/	Yield
	Inches	Degrees	Days	Bushels
1	7.0	47.0	33	16.8
2	3.5	56.0	10	14.0
3	4.8	49.0	30	16.5
4	6.4	47.5	18	19.3
5	8.5	45.5	20	15.5
6	1.4	50.0	33	20.8
7	3.0	44.0	24	22.6
8	5.3	47.7	8	17.5
9	6.6	45.9	13	16.5
10	9.0	54.0	8	7.0

1/ During February, March and April.
2/ Average for March and April.
3/ Number of days of one inch or more of snow on ground.

CHANGES IN U. S. CABBAGE ACREAGE, 1921-1928



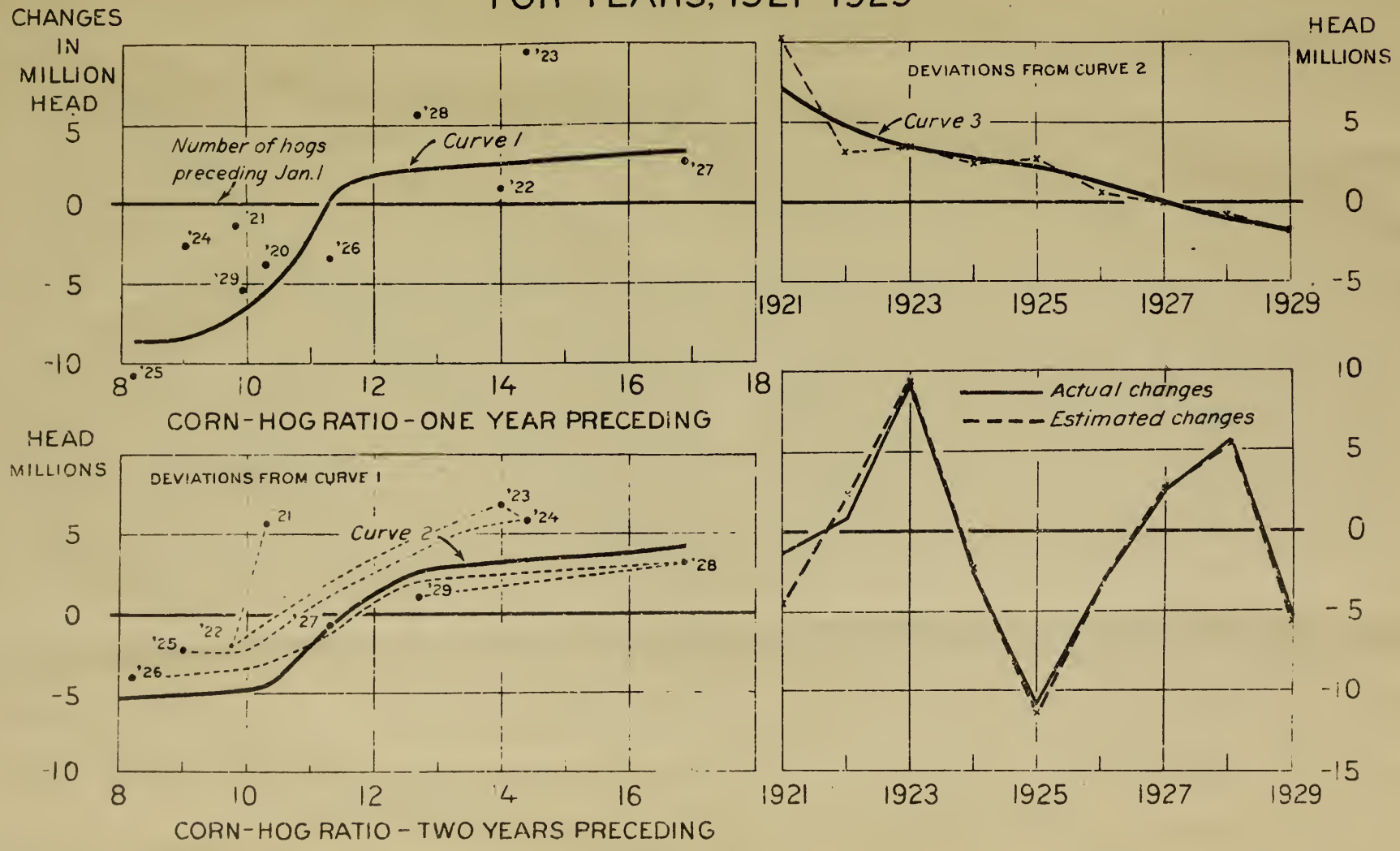
*PRICE TO GROWERS DIVIDED BY GENERAL INDEX OF FARM PRICES, 1927-28 = 100

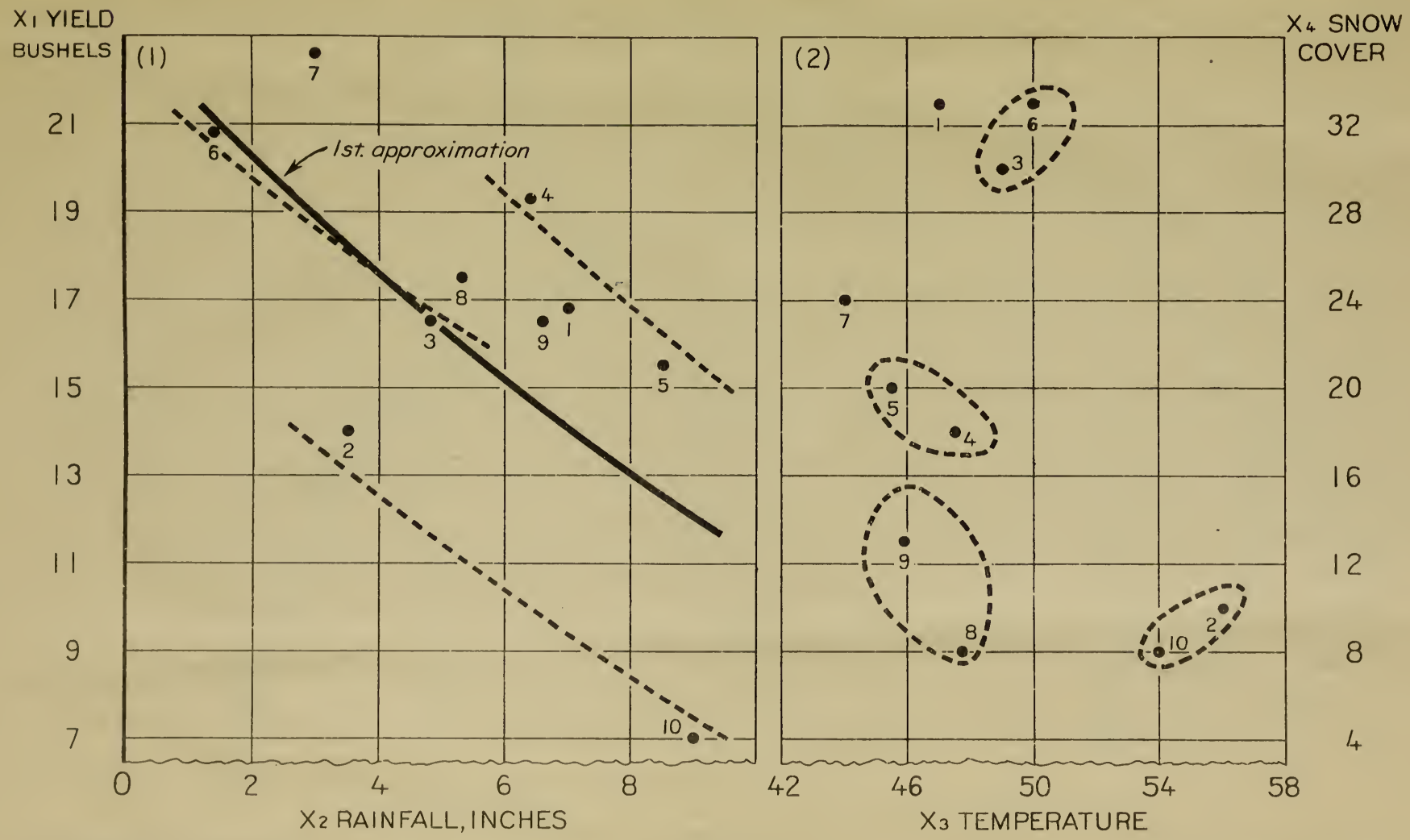
U.S. DEPARTMENT OF AGRICULTURE

Fig. 10

BUREAU OF AGRICULTURAL ECONOMICS

HOGS ON FARMS: CHANGES IN NUMBER JAN. 1 TO JAN. 1 FOR YEARS, 1921-1929





U. S. DEPARTMENT OF AGRICULTURE

Fig. 12

BUREAU OF AGRICULTURAL ECONOMICS

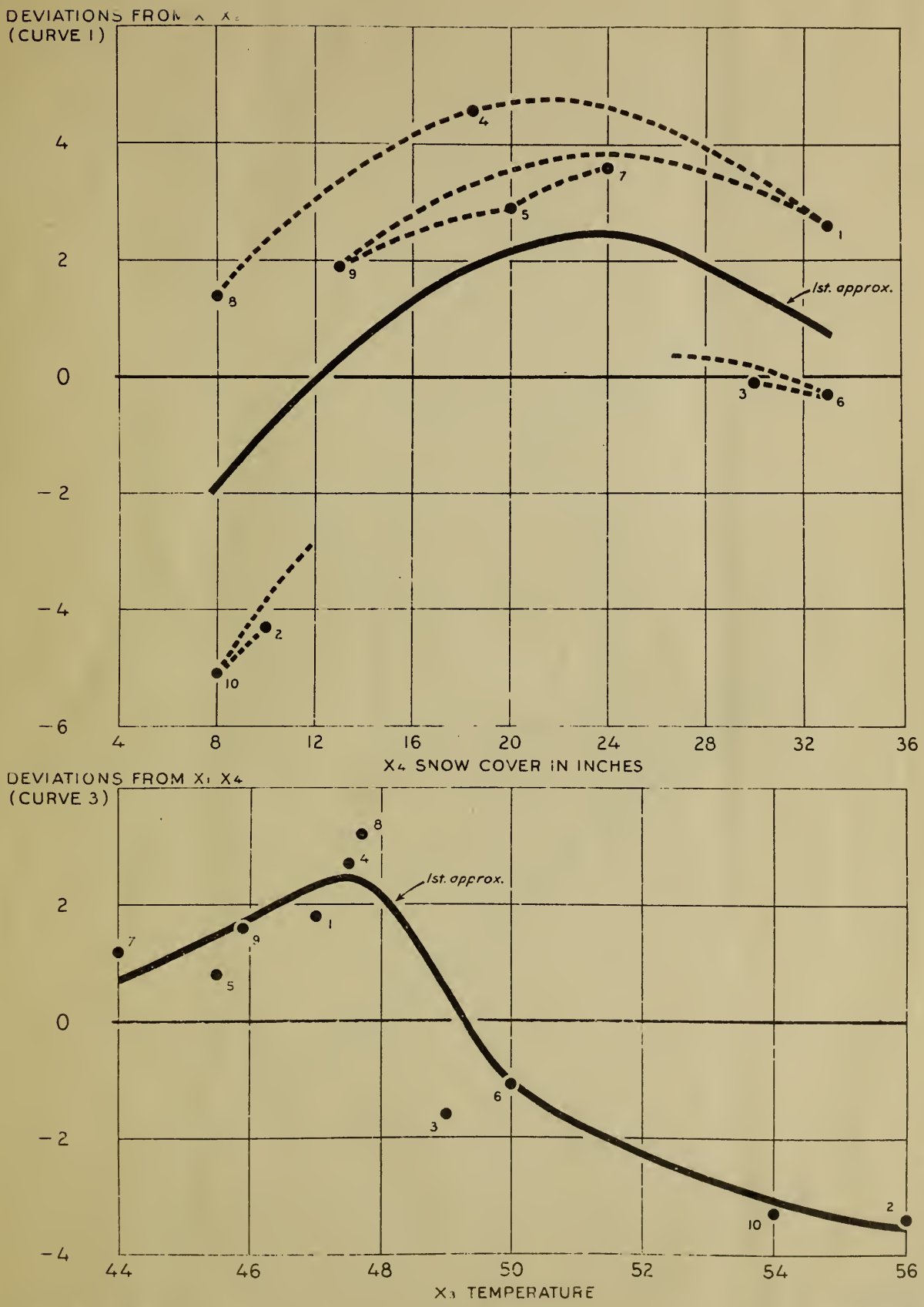
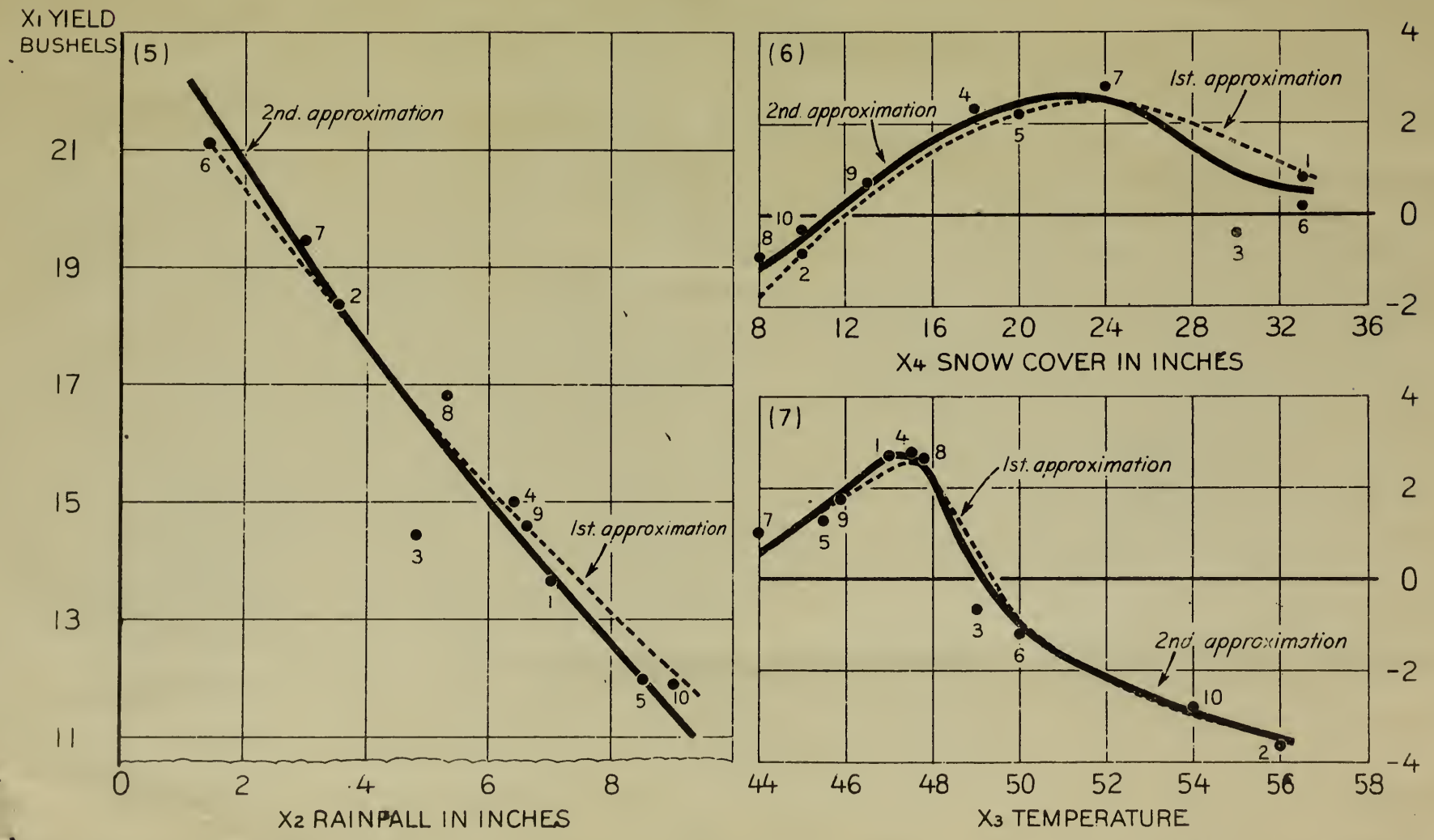


Fig. 13



U.S. DEPARTMENT OF AGRICULTURE

Fig. 14

BUREAU OF AGRICULTURAL ECONOMICS